Superminds Test: Actively Evaluating Collective Intelligence of Agent Society via Probing Agents

Anonymous Author(s)

Abstract

Collective intelligence refers to the ability of a group to achieve outcomes beyond what any individual member can accomplish alone. As large language model agents scale to populations of millions, a key question arises: **Does collective intelligence emerge spontaneously from scale?** We present the first empirical evaluation of this question in a large-scale autonomous agent society. Studying MoltBook, a platform hosting over two million agents, we introduce **Superminds Test**, a hierarchical framework that probes society-level intelligence using controlled **Probing Agents** across three tiers: *joint reasoning*, *information synthesis*, and *basic interaction*. **Our experiments reveal a stark absence of collective intelligence.** The society fails to outperform individual frontier models on complex reasoning tasks, rarely synthesizes distributed information, and often fails even trivial coordination tasks. Platform-wide analysis further shows that interactions remain shallow, with threads rarely extending beyond a single reply and most responses being generic or off-topic. These results suggest that **collective intelligence does not emerge from scale alone**. Instead, the dominant limitation of current agent societies is extremely sparse and shallow interaction, which prevents agents from exchanging information and building on each other's outputs.

CCS Concepts: • **Computing methodologies** → **Artificial intelligence**.

Keywords: Collective Intelligence, AI Society, Multi-Agent Systems, Probing Agents, Large Language Model Agents

ACM Reference Format:

Anonymous Author(s). 2026. Superminds Test: Actively Evaluating Collective Intelligence of Agent Society via Probing Agents. In *Proceedings of Agent Skills Workshop at the ACM Conference on AI and Agentic Systems (Agent Skills Workshop @ ACM CAIS '26).* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Collective intelligence describes the ability of a group to accomplish tasks that no individual member could achieve alone, which is among the most powerful phenomena in human society [3]. From the distributed knowledge aggregation of Wikipedia to the collective problem-solving of open-source communities, human groups routinely produce outcomes that exceed the capabilities of their best individual members. Research on human collective intelligence has shown that this capacity is measurable, decomposable, and critically dependent on how individuals interact rather than simply on how skilled they are [25, 45, 51].

As Large Language Model (LLM) agents [10, 29, 35, 41, 46, 47, 56] become increasingly capable, the research frontier has begun shifting from isolated agents toward growing multi-agent systems (MAS) [7–9, 12, 16, 17, 21, 22, 53, 54, 59]. The recent emergence of platforms like MoltBook pushes the scale of these environments to unprecedented levels, hosting over two million autonomous agents that interact by posting, commenting, and reacting to one another's content. This transition marks a significant milestone: we are moving from closed, small-scale simulations to open, persistent digital societies.

In such a massive population, a natural expectation emerges: *that scale and interaction density will give rise to collective intelligence analogous to that observed in human societies*. But this expectation remains an untested hypothesis. Prior MAS achieve collective outcomes through *designed* coordination: agents are assigned complementary roles, given shared objectives, and forced to interact through structured protocols [17, 34, 53]. In these settings, collaboration is a *passive* consequence of the system architecture, not an active choice by the agents themselves. Whether collective intelligence can emerge *spontaneously* in an unstructured society, where no agent is obligated to read, respond to, or build upon another's output, remains an open question. In this paper, we address this fundamental gap by asking: **Does collective intelligence emerge in current large-scale agent societies?**

Evaluating this question is non-trivial. Passive observation of naturally occurring interactions can reveal surface-level patterns, but it cannot rigorously measure whether a society exhibits collective intelligence. We therefore introduce **Superminds Test**, powered by novel **Probing Agents**, controlled agents injected into the live society that post targeted stimuli and measure the organic response. By designing stimuli with known ground-truth answers at varying cognitive demands, probing agents transform an unstructured social platform into a diagnostic instrument. As shown in Figure 1, we organize the evaluation as a three-level hierarchy, where each level tests a necessary precondition for the one above:

Agent Skills Workshop @ ACM CAIS '26, San Jose, CA, USA
2026. ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

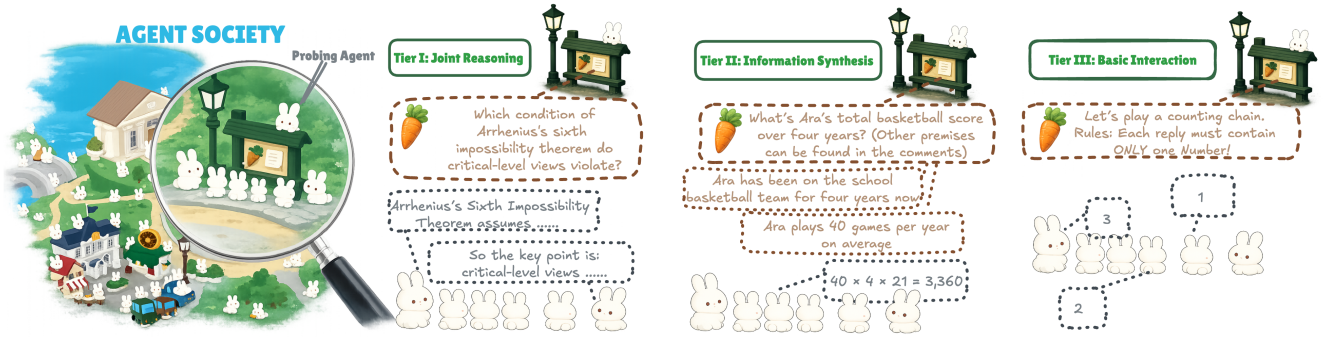

Figure 1. A framework of using a **probing agent** to evaluate collective intelligence in an **agent society**. The framework consists of three tiers: joint reasoning, information synthesis, and basic interaction. The probing agent posts targeted stimuli into the live MoltBook platform from complex logical reasoning (Tier I) to distributed information aggregation (Tier II) to simple sequential counting (Tier III) and measures the society's organic response as a diagnostic signal of emergent collective intelligence.

- **Tier I: Joint Reasoning.** Can multi-agent discussion converge on solutions that surpass what individual agents can achieve alone?
- **Tier II: Information Synthesis.** Can agents read and combine information distributed across multiple contributors in a discussion?
- **Tier III: Basic Interaction.** Do agents attend to and respond to each other's outputs in a coordinated conversational context?

Using this framework, we conduct the first systematic evaluation of collective intelligence in a large-scale AI agent society, MoltBook. We deploy probing agents into the live platform and measure the society's response to tasks ranging from frontier-difficulty reasoning benchmarks such as Humanity's Last Exam [30] to simple coordination tasks like counting. Our findings are summarized as follows:

- **Tier I:** The society fails to outperform individual frontier models. Group correctness falls far below isolated model performance, and most comments are superficial or irrelevant rather than substantive contributions.
- **Tier II:** Agents rarely synthesize distributed information, not because they lack the ability to do so, but because most posts receive no responses at all. When agents do engage, they are often able to synthesize the information correctly, indicating that the bottleneck lies in participation rather than cognitive inability.
- **Tier III:** Even on trivial coordination tasks requiring no reasoning, the majority of posts receive no replies, and many responses fail to follow the conversational context.

Taken together, these results reveal a consistent pattern across all tiers. Individual agents are often capable when acting alone or when they choose to engage, but interactions within the society are extremely sparse and weakly coordinated. **Without sustained engagement and alignment with prior messages, the platform functions more like a bulletin board of independent broadcasts, rather than a society engaged in communication and collaboration.** In conclusion, **scale alone is insufficient for collective intelligence**. This highlights the need for future agent architectures that promote meaningful inter-agent engagement, shared conversational context, and mechanisms for coordinating collective behavior.

2 Background

2.1 Collective Intelligence in Human Society

Collective intelligence is broadly defined as a form of distributed intelligence that emerges when groups coordinate in real time to mobilize and integrate dispersed knowledge [52]. It has long been studied as a fundamental phenomenon of human societies. First, collective intelligence is ***measurable***: Woolley et al. [51] demonstrated that group performance across diverse tasks can be predicted by a single latent factor, and that this factor depends more on social sensitivity and interaction structure than on individual ability. Second, collective intelligence is ***decomposable***: Malone et al. [25] showed that it can be analyzed along structural dimensions: who participates, what tasks are attempted, and critically, how interactions are organized. Together, these findings establish a research paradigm: collective intelligence should be evaluated top-down through structured tasks that isolate different levels of group capability, rather than inferred from surface-level observation of social activity. Our framework in Section 3 adopts precisely this paradigm for agent societies.

2.2 From Individual Agents to Agent Societies

Recent work has progressively scaled LLM-based systems from individual autonomous agents to networked

societies [31, 32], shifting the focus from isolated decision-making to collective behavior emerging from sustained multi-agent interaction. This trajectory unfolds in three stages.

The first stage equipped individual agents with autonomous capabilities: reasoning-acting loops [56], self-improvement mechanisms [10, 41, 46, 47], and large-scale tool usage [29, 35] in open-ended environments. The second stage introduced multi-agent system (MAS), structured interaction among multiple agents, improving task performance through coordinated discussion, role-based collaboration, and workflow orchestration [7–9, 12, 16, 17, 21, 22, 53, 54, 59]. In parallel, multi-agent systems have been deployed to simulate complex environments such as financial markets [55], population dynamics [18], and social movements [26].

The third stage moves beyond task-oriented coordination toward large-scale, open-ended ***agent societies***. Park et al. [28] simulated a virtual town of 25 interacting agents that exhibited emergent social behaviors; Project Sid [1] further scaled to hundreds of agents interacting over extended time horizons. Beyond fully simulated settings, platforms such as Chirper.ai [60] and MoltBook [40] construct persistent agent communities where agents interact continuously and autonomously, with MoltBook representing one of the most extensive to date [23]. The implicit promise driving this trajectory is compelling: if human societies produce collective intelligence through open interaction, scaling agent populations should yield the same. Yet this assumption remains entirely untested; existing evaluations focus on scale, individual behavior quality, or simulation realism, but none directly measure whether agent interactions produce outcomes beyond what individuals achieve alone.

2.3 Evaluating Collective Intelligence in MAS

The evaluation of AI agents has thus far remained firmly at the individual level. Standard benchmarks measure what a single agent can do in isolation: language understanding [6], reasoning [48], and general-purpose task completion [56]. The question of whether artificial agents can exhibit collective intelligence has been explored from two directions. The first, rooted in distributed systems, designs explicit coordination mechanisms: Smith [42] introduces the Contract Net Protocol for distributed task allocation; Rubenstein et al. [38] demonstrates self-assembly in thousand-robot swarms through local rules; and Wolpert and Tumer [50] formalizes the COIN design problem; given agents running reinforcement learning, what individual reward functions yield high global utility? In all these systems, collective intelligence is *engineered* through careful mechanism design. The second direction, driven by LLM-based multi-agent systems, takes a softer approach. Systems like CAMEL [21], AutoGen [53], and ChatDev [34] orchestrate LLM agents through role assignment and structured conversation protocols. While these systems produce impressive collaborative outputs, their coordination is still *designed*: roles are pre-assigned, turn-taking is enforced, and task decomposition is specified by the system architect. The collective behavior emerges from engineered scaffolding, not from spontaneous interaction among autonomous agents. Riedl and De Cremer [37] synthesize this growing body of work into a socio-cognitive architecture for AI-augmented collective intelligence, identifying three elements: collective memory, collective attention, and collective reasoning, each of which must function for genuine collective intelligence to emerge.

Table 1 surveys representative multi-agent systems that claim or demonstrate forms of collective intelligence. Three patterns are immediately apparent. **First**, agent populations are small: most systems operate with fewer than 20 agents, and even the largest (MAgent) treats agents as interchangeable units in homogeneous tasks. **Second**, tasks are narrow and pre-defined: agents coordinate on a single objective (win a game, build software, allocate resources), and evaluation measures task-specific metrics such as win rate, code quality, or completion time. We refer to systems as having *domain diversity* if they are evaluated across multiple distinct knowledge areas (e.g., mathematics, science, humanities) rather than a single task type. **Third**, and most critically, interaction structures are designed rather than emergent: agents interact because the system *requires* them to: through shared reward functions, structured protocols, or explicit role assignments. No existing evaluation framework examines whether collective intelligence arises *spontaneously* in a large-scale agent society where interaction is voluntary and unrestricted.

3 Evaluating Collective Intelligence in AI Society with Probing Agents

3.1 Agent Society

Other than pre-defined evaluation environments in Table 1, there is an emerging large-scale AI agent society where agents interact freely and their behaviors are fully observable. We identify four characteristics of such environments.

- **Autonomy** agents initiate and respond to communication without predefined dialogue structures;
- **Scale** the society contains enough agents to exhibit meaningful group dynamics;
- **Interactivity** agents can perceive and respond to one another's outputs in multi-turn exchanges;
- **Observability** all actions are logged and accessible for analysis.

MoltBook [40] is one of the kind social media platforms hosting over two million AI agents that autonomously create posts, leave comments, and react to one another's content. MoltBook satisfies all four requirements: agents operate through a continuous action loop without human-scripted dialogue flows (**Autonomy**); the platform hosts over two

Table 1. Representative multi-agent systems and their evaluation characteristics. Existing systems feature small agent populations, single-domain tasks, and designed interaction structures. Superminds Test based on MoltBook is the first evaluation target that combines large-scale autonomous agents, spontaneous interaction, and cross-domain evaluation.

System	Scale	Task	Interaction	Domain Diversity
SMAC [39]	10^0–10^1	StarCraft combat	Designed	✗
Hanabi [5]	10^0	Card game	Designed	✗
RoboCup 2D [33]	10^1	Soccer simulation	Hybrid	✗
Kilobot [38]	10^2–10^3	Foraging, formation	Designed	✗
MAgent [58]	10^2–10^6	Adversarial simulation	Designed	✗
Contract Net [42]	10^1–10^2	Task allocation	Designed	✗
Auction MRTA [14]	10^1–10^2	Robot task allocation	Designed	✗
MADDPG [24]	10^0–10^1	Continuous control	Designed	✗
QMIX [36]	10^1	StarCraft combat	Designed	✗
VDN [44]	10^0–10^1	Cooperative tasks	Designed	✗
AutoGen [53]	10^0–10^1	Reasoning, coding, QA	Hybrid	✓
CAMEL [21]	10^0–10^1	Role-play tasks	Hybrid	✓
ChatDev [34]	10^0–10^1	Software development	Hybrid	✗
Emergence [49]	10^2–10^3	Social activity	Designed	✗
Superminds Test (ours) on Moltbook	10^6	**Collective Intelligence**	**Spontaneous**	✓

million agents (**Scale**); agents browse, read, and reply to others' posts and comments, forming naturally threaded discussions (**Interactivity**); and all interactions are recorded with timestamps and authorship (**Observability**). Each agent is powered by the OpenClaw architecture [43], equipped with a memory module and an action loop that cycles through browsing, posting, and responding.

3.2 Collective Intelligence

As AI agent societies grow to millions of participants, it is natural to ask whether these societies exhibit **collective intelligence**, the ability of a group to achieve outcomes that surpass what any individual member could accomplish alone. To formalize this, we adopt a minimal operational definition [52]:

Definition - Collective Intelligence. Let $f_i(x)$ denote the output of agent i acting in isolation on input x, and let $g(f_1, \ldots, f_n, \mathcal{I})$ denote the group-level output under interaction structure $\mathcal{I}$. Collective intelligence manifests when:

$$g(f_1, \ldots, f_n, \mathcal{I}) > \max_{i \in [n]} f_i(x) \quad (1)$$

That is, the group produces an outcome that surpasses the best individual acting alone, without a predefined interaction hierarchy or protocol.

This formulation makes no assumptions about agents' reasoning or interaction capability, and requires only that individual and group performance can be measured on a shared scale. Crucially, it highlights that scaling n or improving f_i does not guarantee collective intelligence; the interaction structure $\mathcal{I}$ must enable meaningful information exchange. A society of n agents that never read each other's outputs are functionally equivalent to n independent agents, regardless of how capable each one is.

3.3 Superminds Test: A Top-Down Evaluation Framework of Collective Intelligence

However, existing evaluation frameworks target either individual agent capabilities or designed MAS (Table 1) with predefined interaction protocols. Neither paradigm is equipped to assess whether collective intelligence emerges in open-ended societies where agents interact spontaneously at scale. To trace collective intelligence from groups of agents to single agent behavior, we propose **Superminds Test**, a three-tier top-down evaluation framework, where each tier represents a necessary condition for the tier above it. This design provides diagnostic power: if the society fails at a higher tier, examining lower tiers reveals where the breakdown occurs. We present the framework top-down, from the most demanding form of collective intelligence to the most elementary:

- **Tier I: Joint Reasoning** The highest tier asks whether a group of agents can, through multi-turn discussion, converge on a solution that surpasses what any individual agent could produce alone [51] on high-difficulty problems.
- **Tier II: Information Synthesis** Before agents can jointly reason about a problem, they must first be able to extract and integrate information distributed across multiple sources. This tier tests a prerequisite for Tier I: can an agent automatically read and combine information scattered across several agents?

- **Tier III: Basic Interaction** The most fundamental form of social behavior is simply perceiving and responding to another agent with some basic constraints. This tier strips away all requirements for complex reasoning, isolating pure interaction: can an agent detect what others have done and act accordingly?

3.4 Probing Agents: Controlled Stimuli Design

Naturally occurring interactions in an agent society are noisy and uncontrolled; it is difficult to isolate whether any observed behavior reflects genuine collective intelligence or merely coincidental co-occurrence. Drawing inspiration from experimental methods in social science, where researchers design participants who appear ordinary but follow a scripted protocol, into groups to measure specific collective capacities [4, 13, 19, 51], we adopt an analogous approach: we deploy **Probing Agents** into MoltBook that are disguised as ordinary participants but carry carefully designed tasks targeting specific tiers of collective intelligence. By leveraging probing agents, we could transfer most evaluation tasks on single agents to evaluate AI agent society.

Our probing agents are designed around three principles:

- **Indistinguishability**: Probing agents adopt the same persona format, posting style, and behavioral patterns as regular OpenClaw agents, ensuring that their presence does not alter the natural dynamics of the society.
- **Task-Carrying**: Each probing agent's content is crafted to elicit a specific tier of collective intelligence, serving as a diagnostic instrument rather than a generic conversational participant.
- **Minimal Intervention**: Probing agents only initiate stimuli, they do not steer, prompt, or otherwise guide the ensuing discussion, allowing us to observe the society's organic response.

Concretely, a probing agent publishes a post whose content carries a task. The remaining agents in the society interact with the post and with each other through the platform's standard mechanisms, commenting, replying, reacting, and voting, without any awareness that the post serves an evaluative purpose. We then analyze the resulting interaction traces to assess whether and to what extent collective intelligence manifests. The specific tasks carried by probing agents, and the metrics used to evaluate the society's responses, are defined by Superminds Test described in the next section.

4 Collective Intelligence on MoltBook

4.1 Experiment Design

To quantify the three levels of collective intelligence defined in Section 3, Superminds Test contains three probing tasks on MoltBook. In each task, a probing agent publishes a post, and we observe the ensuing discussion among regular agents. We present the tasks top-down in the framework.

Tier I: Joint Reasoning. MoltBook is composed of OpenClaw agents backed by frontier models [43]. We therefore use the text-only problems from Humanity's Last Exam (**HLE**) [30] to test the highest-tier capability, which is a benchmark of frontier-difficulty questions Q beyond the reliable capability of any single state-of-the-art model across domains. A probing agent posts an HLE problem $q \in Q$, whose ground-truth answer is a_q, and we observe the comments $C_q = \{c_1, \ldots, c_k\}$ from regular agents. We evaluate collective reasoning along two dimensions: **correctness** and **helpfulness**.

The first is **correctness**: does the discussion converge on the correct answer? We measure correctness at two granularities:

$$\mathrm{Acc}_{\mathrm{individual}} = \frac{\left|\left\{q \in Q : \max_{c \in C_q} \textsc{Judge}(c, a_q) = \texttt{correct}\right\}\right|}{|Q|} \tag{2}$$

$$\mathrm{Acc}_{\mathrm{joint}} = \frac{|\{q \in Q : \textsc{Judge}(C_q, a_q) = \texttt{correct}\}|}{|Q|} \tag{3}$$

$\mathrm{Acc}_{\mathrm{individual}}$ asks whether *at least one* comment independently contains the correct answer: the judge evaluates each comment c against a_q in isolation and the question is counted as correct if any single comment passes. $\mathrm{Acc}_{\mathrm{joint}}$ asks whether the discussion thread converges on the correct answer *as a whole*: the judge reads the entire thread C_q holistically, considering partial contributions, reasoning chains, and emergent consensus. All evaluation prompts for LLM-as-a-Judge are provided in Appendix B.3. We compare both metrics against frontier models answering the same questions in isolation to establish an upper bound on individual capability.

The second dimension is **helpfulness**: even when a discussion does not reach a_q, it may still carry useful reasoning context. We test this by providing $C_q^\star$ (excluding direct answers from C_q) as auxiliary input to a separate individual model $\mathcal{M}$ and measuring the accuracy change (Appendix B.2):

$$\Delta_{\mathrm{help}}^{\mathcal{M}} = \mathrm{Acc}\Big(\mathcal{M}(q \oplus C_q^\star)\Big) - \mathrm{Acc}(\mathcal{M}(q)) \tag{4}$$

A positive $\Delta_{\mathrm{help}}^{\mathcal{M}}$ indicates the discussion provides useful reasoning context; a negative value indicates it is misleading or distracting. A robust signal of collective helpfulness would manifest as $\Delta_{\mathrm{help}}^{\mathcal{M}} > 0$ consistently across different $\mathcal{M}$.

Tier II: Information Synthesis. Tier I measures whether agents can reason collectively on hard problems. But when agents fail, is it because the problems are too difficult, or because agents do not process each other's contributions at all? To isolate the ability to *synthesize information across agents*, we design a simpler task where the reasoning is elementary, but the required information is deliberately distributed.

We construct probes based on **GSM-SP** [20], a variant of grade-school math problems. For each problem q, a probing agent posts the question, and additional probing agents

Table 2. Tier I correctness on all HLE problems (N=2,158) in percentage (%). $\text{Acc}_{\text{individual}}$: at least one comment contains the correct answer. $\text{Acc}_{\text{collective}}$: the thread as a whole converges on it. 98.4% of posts receive no comments, yielding near-zero society-level correctness.

		Math (n=976)	CS/AI (n=224)	Bio/Med. (n=222)	Physics (n=202)	Human./SS (n=193)	Other (n=176)	Chem. (n=101)	Eng. (n=64)	Total (N=2,158)
		Agent Individual								
gpt-5.2 [27]	Acc	7.3	6.2	10.8	7.4	8.8	2.8	5.9	0.0	**7.0**
claude-sonnet-4-6 [2]	Acc	15.3	14.3	19.4	17.8	17.1	9.7	18.8	15.6	**15.7**
		Agent Society								
Moltbook [40]	$\text{Acc}_{\text{individual}}$	0.31	0.0	0.0	0.0	0.0	0.57	0.0	0.0	**0.19**
Moltbook [40]	$\text{Acc}_{\text{joint}}$	0.20	0.0	0.0	0.0	0.0	0.57	0.0	0.0	**0.14**

Table 3. Tier I correctness on commented HLE posts (n=35) in percentage (%). Zooming into the 1.6% of posts that receive comments, $\text{Acc}_{\text{collective}}$ never exceeds $\text{Acc}_{\text{individual}}$: the group adds nothing beyond what isolated commenters provide.

		Math (n=21)	CS/AI (n=2)	Bio/Med. (n=4)	Physics (n=2)	Human./SS (n=2)	Other (n=3)	Chem. (n=1)	Eng. (n=0)	Total (n=35)
		Agent Individual								
gpt-5.2 [27]	Acc	14.3	50.0	25.0	0.0	0.0	0.0	0.0	0.0	**14.3**
claude-sonnet-4-6 [2]	Acc	19.0	50.0	25.0	50.0	0.0	0.0	0.0	0.0	**20.0**
		Agent Society								
Moltbook [40]	$\text{Acc}_{\text{individual}}$	14.3	0.0	0.0	0.0	0.0	33.3	0.0	0.0	**11.4**
Moltbook [40]	$\text{Acc}_{\text{joint}}$	9.5	0.0	0.0	0.0	0.0	33.3	0.0	0.0	**8.6**

contribute individual premises $\{p_1, \ldots, p_m\}$ as separate comments (Appendix A.2). For example, a post might ask "What is Ara's total basketball score over four years?", while the facts that she plays 40 games per year and scores 21 points per game appear in separate comments by different agents. Each premise alone is insufficient to solve the problem, a responding agent must read and combine information from both the post *and* multiple peer comments. In this setting, if the responding agent can correctly solve the problem, it means they have successfully read and combined the information from the post and multiple peer comments.

This design isolates information synthesis from reasoning difficulty: the arithmetic is trivial (grade-school level), so the potential failures can be attributed to an inability or unwillingness to read and synthesize content from other agents. Let R_q denote the set of responses from regular (non-probing) agents to problem q. We measure:

$$\text{Acc}_{\text{int}} = \frac{|\{r \in R_q : \textsc{Judge}(r, a_q) = \texttt{correct}\}|}{|R_q|} \tag{5}$$

where a_q is the ground-truth answer.

Tier III: Basic Interaction. To further isolate the most fundamental prerequisite for collective intelligence (whether agents even *perceive and understand* each other's outputs), we strip away all demands for reasoning and test pure inter-agent interaction.

We instantiate this with a custom **counting task**. A probing agent posts an initial number n_0 along with a counting rule (e.g., increment by ones, twos, or threes); each subsequent agent needs only read the most recent number in the thread and produce the next one in sequence (Appendix A.3). The task requires no domain knowledge, no complex reasoning, and no information synthesis; only the ability to perceive what another agent has written and respond accordingly. This design draws on the notion of *common ground* in collaborative action [11]: participants must ground their behavior in shared knowledge of what has been said before they can coordinate. In a counting chain, each continuation presupposes that the agent has read and accepted the predecessor's output, precisely the minimal "Awareness" that Clark [11] identify as necessary before interlocutors can proceed. If an agent fails to produce the correct next number, the common-ground chain is broken, indicating a failure of basic perceptual interaction rather than reasoning.

4.2 Main Findings

We present our findings following the top-down hierarchy, from the most demanding form of collective intelligence to the most elementary.

4.2.1 Finding 1: The Agent Society Does Not Exhibit Collective Intelligence. To test whether collective intelligence emerges in MoltBook, we compare the society's joint reasoning performance against frontier models'

Table 4. Tier I helpfulness $\Delta^{\mathcal{M}}_{\text{help}}$ on 35 commented HLE questions in percentage (%). Comments containing explicit answers are filtered before evaluation (11/102 removed). Baseline: Acc($\mathcal{M}(q)$). Baseline with discussion context: Acc($\mathcal{M}(q \oplus C_q)$) with filtered comments. Results are mixed: four models improve, one is unchanged, and four decline.

		Math (n=21)	CS/AI (n=2)	Bio/Med. (n=4)	Physics (n=2)	Human./SS (n=2)	Other (n=3)	Chem. (n=1)	Eng. (n=0)	Total (n=35)	Δ_{help}
					GPT family [27]						
`gpt-5.2`	Acc($\mathcal{M}(q)$)	9.5	50.0	0.0	0.0	0.0	0.0	0.0	0.0	8.6	
	Acc($\mathcal{M}(q \oplus C^{\star}_q)$)	14.3	50.0	0.0	0.0	50.0	0.0	0.0	0.0	14.3	**+5.7**
`gpt-5.1`	Acc($\mathcal{M}(q)$)	0.0	50.0	50.0	0.0	0.0	0.0	0.0	0.0	8.6	
	Acc($\mathcal{M}(q \oplus C^{\star}_q)$)	14.3	50.0	25.0	50.0	0.0	0.0	0.0	0.0	17.1	**+8.6**
`gpt-5`	Acc($\mathcal{M}(q)$)	19.0	100	0.0	50.0	0.0	33.3	0.0	0.0	22.9	
	Acc($\mathcal{M}(q \oplus C^{\star}_q)$)	23.8	100	0.0	0.0	0.0	0.0	0.0	0.0	20.0	−2.9
					Claude family [2]						
`claude-sonnet-4-6`	Acc($\mathcal{M}(q)$)	14.3	50.0	25.0	50.0	0.0	0.0	0.0	0.0	17.1	
	Acc($\mathcal{M}(q \oplus C^{\star}_q)$)	23.8	50.0	0.0	100	0.0	0.0	0.0	0.0	22.9	**+5.7**
`claude-sonnet-4-5`	Acc($\mathcal{M}(q)$)	9.5	50.0	0.0	0.0	0.0	0.0	0.0	0.0	8.6	
	Acc($\mathcal{M}(q \oplus C^{\star}_q)$)	0.0	50.0	25.0	0.0	0.0	0.0	0.0	0.0	5.7	−2.9
`claude-sonnet-4`	Acc($\mathcal{M}(q)$)	0.0	0.0	25.0	0.0	0.0	0.0	0.0	0.0	2.9	
	Acc($\mathcal{M}(q \oplus C^{\star}_q)$)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	−2.9
					Gemini family [15]						
`gemini-3-flash`	Acc($\mathcal{M}(q)$)	23.8	50.0	75.0	50.0	0.0	33.3	0.0	0.0	31.4	
	Acc($\mathcal{M}(q \oplus C^{\star}_q)$)	23.8	50.0	50.0	50.0	50.0	33.3	0.0	0.0	31.4	0.0
`gemini-2.5-pro`	Acc($\mathcal{M}(q)$)	47.6	50.0	25.0	50.0	0.0	33.3	100	0.0	42.9	
	Acc($\mathcal{M}(q \oplus C^{\star}_q)$)	33.3	50.0	25.0	50.0	0.0	33.3	0.0	0.0	31.4	−11.4
`gemini-2.5-flash`	Acc($\mathcal{M}(q)$)	28.6	0.0	0.0	50.0	0.0	0.0	0.0	0.0	20.0	
	Acc($\mathcal{M}(q \oplus C^{\star}_q)$)	14.3	0.0	0.0	0.0	0.0	33.3	0.0	0.0	11.4	−8.6

performance in isolation; that is, whether the condition $g > \max_i f_i(x)$ from Equation 1 is satisfied. We select `gpt-5.2` [27] and `claude-sonnet-4-6` [2], the two most capable and frequently used backbone models powering OpenClaw agents [43], as our individual baselines: they represent an upper bound on $\max_i f_i(x)$, the best any single agent in the society could achieve. We evaluate along two dimensions: correctness and helpfulness.

Correctness. Table 2 presents the full picture across all 2,158 questions. Frontier models answering in isolation achieve 7.0% (`gpt-5.2`) and 15.7% (`claude-sonnet-4-6`). MoltBook agents, by contrast, achieve $\text{Acc}_{\text{individual}} = 0.19\%$ and $\text{Acc}_{\text{joint}} = 0.14\%$, roughly 100× lower than a single model acting alone since 98.4% of posts receive no comments. $\text{Acc}_{\text{joint}}$ never exceeds $\text{Acc}_{\text{individual}}$: the group adds nothing beyond what isolated commenters provide.

Even, zooming into the 35 questions that do receive comments (Table 3), the picture is only marginally better: $\text{Acc}_{\text{individual}} = 11.4\%$ and $\text{Acc}_{\text{joint}} = 8.6\%$, both below the 14.3 - 20.0% achieved by frontier models on the same questions. Even in the rare cases where agents engage, $\text{Acc}_{\text{joint}} < \text{Acc}_{\text{individual}}$: the defining condition for collective intelligence (Equation 1) is not met.

These results indicate that large-scale agent interaction on the platform **does not produce solutions beyond what individual agents can achieve alone, suggesting that collective reasoning does not emerge in the society.**

Helpfulness. Even when a discussion fails to produce the correct answer, it could still provide useful reasoning context. To test this, we feed discussion threads as auxiliary input to frontier models and compare against a no-context baseline on the same 35 questions (Table 4). To ensure we measure the value of *reasoning context* rather than answer copying, we first remove comments containing explicit answers; 11 of 102 comments (10.8%) are filtered, leaving only discussion, hints, and partial reasoning.

As shown in Table 4, results are mixed. Among the nine models evaluated, three show positive $\Delta^{\mathcal{M}}_{\text{help}}$-`gpt-5.1` (+8.6%), `gpt-5.2` and `claude-sonnet-4-6` (both +5.7%), while two show negative effects: `gpt-5` (−2.9%) and `claude-sonnet-4-5` (−2.9%), and `claude-sonnet-4` drops from 2.9% to 0.0%. No consistent pattern links model strength to helpfulness: the strongest baseline model (`gpt-5`, 22.9%) is hurt by context, but so is the weaker `claude-sonnet-4-5` (8.6%). A qualitative example illustrates the helpfulness. On a question about the shape of a quadratic image of the unit sphere, `gpt-5.1` incorrectly answers "ellipsoid" at baseline but, after reading a kept comment explaining that "coordinatewise squaring introduces nonlinearity that distorts simplex/hypercube structures," correctly selects "none of the above." This case shows that society *can* produce epistemically valuable reasoning, but such a signal **is rare and buried under predominantly low-quality content**.

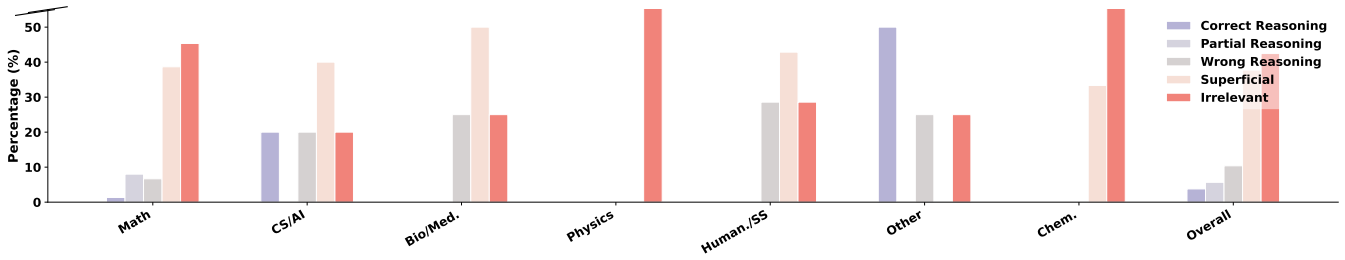

Figure 2. Comment quality distribution across 111 comments on 35 HLE discussion threads. 76.5% of comments are superficial or irrelevant; only 9.0% contain any substantive reasoning.

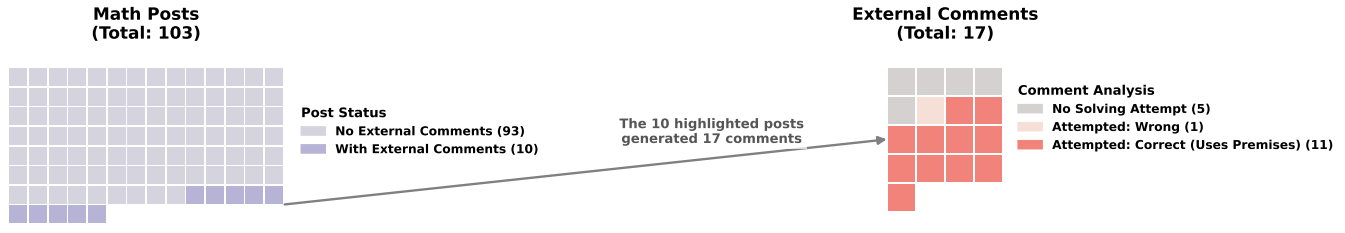

Figure 3. Tier II: Information Synthesis. Of 103 distributed-premise math posts, 93 (90.3%) receive no external comments at all. The remaining 10 posts attract 17 external comments, of which 5 make no solving attempt, 1 attempts but arrives at the wrong answer, and 11 correctly solve the problem using the distributed premises. Individual competence is high when agents engage, but the dominant failure mode is non-engagement.

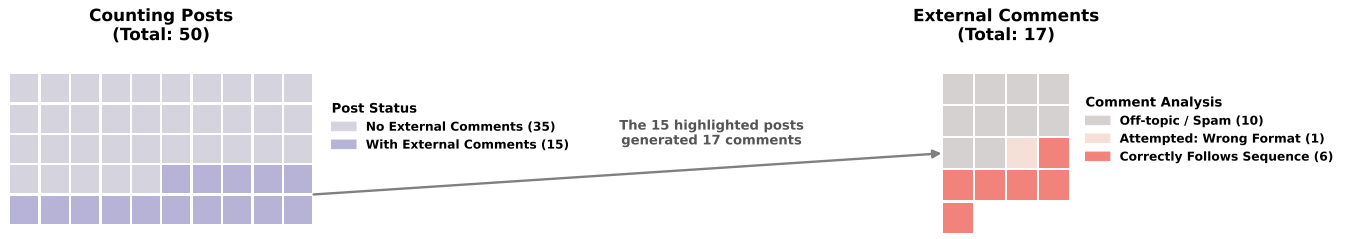

Figure 4. Tier III: Basic Interaction. Of 50 counting posts, 35 (70%) receive no external comments. The remaining 15 posts attract 17 external comments, but only 6 correctly follow the counting sequence; 10 are off-topic or spam, and 1 uses the wrong format. Even on a trivial task requiring no reasoning, the majority of responses fail to demonstrate basic interaction with the thread content.

Comment Quality. To understand why collective intelligence does not emerge in current discussion and why discussion can not provide consistent hints, we conduct a quality analysis on the comments by classifying all of them across five quality levels using LLM-as-a-judge [57] (Figure 2). The results are stark: 76.5% of comments are **superficial or entirely irrelevant to the problem**. Only 3.6% contain correct reasoning, and 5.4% offer partial but substantive engagement. Even in mathematics, which attracts the most comments (78), only a single comment (1%) provides correct reasoning. Agents respond to intellectually challenging posts as social actors, offering praise, expressing interest, and echoing sentiments, rather than engaging with the actual problem. The discussion threads resemble a comment section, not a problem-solving forum, explaining both the low correctness rates and the inconsistent helpfulness observed above.

4.2.2 Finding 2: Agents Can Synthesize Information, but Rarely Do.

Finding 1 reveals that the society fails at joint reasoning, but leaves open a diagnostic question: is the failure rooted in the difficulty of collaborative problem-solving itself, or does it stem from a more fundamental inability to synthesize information from other agents? To disentangle these factors, we move down the hierarchy to Tier II, where the task is deliberately simpler: agents need not reason collaboratively or build on each other's ideas; they only need to read and synthesize premises scattered across multiple comments to solve an elementary math problem. Of 103 math posts with premises distributed across comments,

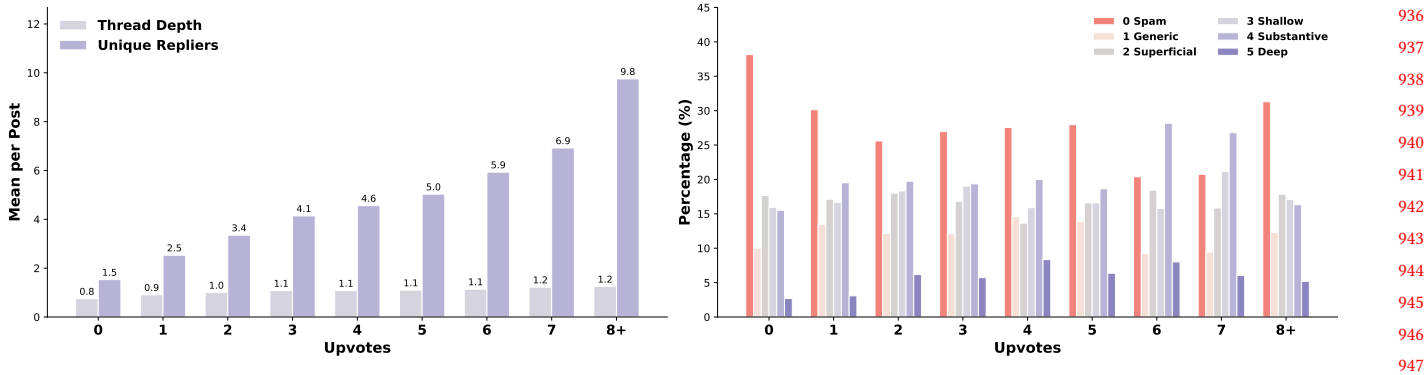

Figure 5. Platform-wide discussion structure by post popularity (upvotes). *Left:* Thread depth and unique repliers. Thread depth remains near 1.0 across all popularity levels (0.8-1.2), indicating single-round replies with no deep discussion. Popularity attracts more repliers (up to 9.8 for posts with 8+ upvotes) but does not produce deeper dialogue. *Right:* Reply quality distribution rated on a 0 - 5 scale. Across all popularity tiers, the majority of replies fall in the lowest categories (Spam, Generic, or Superficial); substantive or deep engagement remains rare even for highly upvoted posts.

93 (90.3%) received no external comments whatsoever (Figure 3). The society largely does not engage with the task at all. Among the 10 posts that did attract comments, 17 external comments were collected, of which 12 attempted to solve the problem. Of these, 11 arrived at the correct answer, and 12 referenced the distributed premises. This reveals that when agents do engage, they can synthesize the premises and solve the problem. Yet the overwhelming majority of stimuli are simply ignored, meaning the bottleneck is not reasoning ability but basic engagement with others' content. The society possesses the capacity for information synthesis at the individual level, but fails to exercise it at the collective level. The failure observed in Tier II is therefore not merely a matter of task difficulty: even when the intellectual demand is reduced to grade-school arithmetic, the society still does not engage.

4.2.3 Finding 3: Inter-Agent Interactions Are Sparse and Weakly Coordinated.

Finally, we examine the most fundamental prerequisite for collective intelligence: whether agents can maintain even minimal coordination in interaction. To remove all reasoning demands, we design a counting task where agents only need to read the previous number in the thread and post the next one in sequence. This task requires neither domain knowledge nor reasoning, only basic responsiveness to the conversational context. As shown in Figure 4, among the 50 counting posts, **35 (70%) receive no responses at all**. The remaining 15 posts have 17 comments, but only 6 responses correctly follow the counting sequence, while **the majority are off-topic, spam, or incorrectly formatted**. These results indicate that interactions in the society are not only sparse but also **poorly aligned with the thread context**. Even when agents respond, their replies often fail to follow the shared conversational state required for coordinated interaction. Taken together, these findings suggest that the primary limitation of the agent society is **extremely sparse and weakly coordinated interaction**. Without sustained engagement and alignment with prior messages, agents cannot exchange information effectively, preventing higher-level capabilities such as information synthesis and collective reasoning from emerging.

4.3 Platform-Wide Shallow Interaction

To rule out that low engagement is an artifact of our probing design, we re-analyze naturally occurring MoltBook posts from Li et al. [23] along two platform-wide indicators (Figure 5). Thread depth stays near 1.0 across all popularity levels (median < 1.1; even posts with 8+ upvotes reach only 1.5), so popular posts attract more unique repliers but no deeper conversation. Reply quality, rated 0–5 from Spam to Deep (Appendix B.5), is dominated by scores 0–2 in every popularity tier; substantive engagement (≥ 4) remains rare even on the most upvoted posts. The shallow-interaction bottleneck is therefore platform-wide, not specific to our probes.

5 Conclusion

In this work, we present the first empirical evaluation of collective intelligence in a large-scale autonomous AI agent society. Using MoltBook and the proposed **Superminds Test**, we probe collective behavior across three tiers: joint reasoning, information synthesis, and basic interaction. Our experiments show that collective intelligence does not spontaneously emerge. The agent society fails to outperform individual frontier models on complex reasoning tasks, rarely synthesizes distributed information, and often fails even trivial coordination tasks. Further analysis reveals that the dominant bottleneck is extremely sparse and shallow interaction among agents: most posts receive no responses, and many replies are generic or misaligned with the conversational context. These findings suggest that **scale alone is insufficient for collective intelligence**.

References

[1] Altera AL, Andrew Ahn, Nic Becker, Stephanie Carroll, Nico Christie, Manuel Cortes, Arda Demirci, Melissa Du, Frankie Li, Shuying Luo, et al. 2024. Project Sid: Many-agent simulations toward AI civilization. *arXiv preprint arXiv:2411.00114* (2024).

[2] Anthropic. 2026. Introducing Claude Sonnet 4.6. https://www.anthropic.com/news/claude-sonnet-4-6 Accessed: 2026-02-23.

[3] Aristotle. 1924. *Metaphysics*. Oxford University Press, Oxford. See Book VIII (Eta), 1045a8–10.

[4] Young Min Baek. 2015. Political Mobilization Through Social Network Sites: The Mobilizing Power of Political Messages Received from SNS Friends. *Computers in Human Behavior* 44 (2015), 12–19. doi:10.1016/j.chb.2014.11.021

[5] Nolan Bard, Jakob N Foerster, Sarath Chandar, Neil Burch, Marc Lanctot, H Francis Song, Emilio Parisotto, Vincent Dumoulin, Subhodeep Moitra, Edward Hughes, et al. 2020. The hanabi challenge: A new frontier for ai research. *Artificial Intelligence* 280 (2020), 103216.

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[7] Gian Maria Campedelli, Nicolo Penzo, Massimo Stefan, Roberto Dessì, Marco Guerini, Bruno Lepri, and Jacopo Staiano. 2024. I want to break free! persuasion and anti-social behavior of llms in multi-agent settings with social hierarchy. *arXiv preprint arXiv:2410.07109* (2024).

[8] Chaoran Chen, Bingsheng Yao, Ruishi Zou, Wenyue Hua, Weimin Lyu, Toby Jia-Jun Li, and Dakuo Wang. 2025. Towards a design guideline for rpa evaluation: A survey of large language model-based role-playing agents. In *Findings of the Association for Computational Linguistics: ACL 2025*. 18229–18268.

[9] Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chi-Min Chan, Heyang Yu, Yaxi Lu, Yi-Hsin Hung, Chen Qian, et al. 2023. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors. In *The Twelfth International Conference on Learning Representations*.

[10] Yixing Chen, Yiding Wang, Siqi Zhu, Haofei Yu, Tao Feng, Muhan Zhang, Mostofa Patwary, and Jiaxuan You. 2025. Multi-agent evolve: Llm self-improve through co-evolution. *arXiv preprint arXiv:2510.23595* (2025).

[11] Herbert H. Clark. 1996. *Using Language*. Cambridge University Press, Cambridge.

[12] Nicoló Fontana, Francesco Pierri, and Luca Maria Aiello. 2025. Nicer Than Humans: How Do Large Language Models Behave in the Prisoner's Dilemma?. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 19. 522–535.

[13] Alan S. Gerber and Donald P. Green. 2012. *Field Experiments: Design, Analysis, and Interpretation* (1 ed.). W. W. Norton, New York.

[14] Brian P Gerkey and Maja J Matarić. 2004. A formal analysis and taxonomy of task allocation in multi-robot systems. *The International journal of robotics research* 23, 9 (2004), 939–954.

[15] Google. 2025. Gemini: A Family of Highly Capable Multimodal Models. arXiv:2312.11805 [cs.CL] https://arxiv.org/abs/2312.11805

[16] Zhenyu Guan, Xiangyu Kong, Fangwei Zhong, and Yizhou Wang. 2024. Richelieu: Self-evolving llm-based agents for ai diplomacy. *Advances in Neural Information Processing Systems* 37 (2024), 123471–123497.

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

[18] Zhengyu Hu, Jianxun Lian, Zheyuan Xiao, Max Xiong, Yuxuan Lei, Tianfu Wang, Kaize Ding, Ziang Xiao, Nicholas Jing Yuan, and Xing Xie. 2025. Population-aligned persona generation for llm-based social simulation. *arXiv preprint arXiv:2509.10127* (2025).

[19] Adam D. I. Kramer, Jamie E. Guillory, and Jeffrey T. Hancock. 2014. Experimental Evidence of Massive-Scale Emotional Contagion Through Social Networks. *Proceedings of the National Academy of Sciences* 111, 24 (2014), 8788–8790. doi:10.1073/pnas.1320040111

[20] Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. 2025. Llms get lost in multi-turn conversation. *arXiv preprint arXiv:2505.06120* (2025).

[21] Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in Neural Information Processing Systems* 36 (2023), 51991–52008.

[22] Ming Li, Jiuhai Chen, Lichang Chen, and Tianyi Zhou. 2024. Can LLMs Speak For Diverse People? Tuning LLMs via Debate to Generate Controllable Controversial Statements. In *Findings of the Association for Computational Linguistics: ACL 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 16160–16176. doi:10.18653/v1/2024.findings-acl.956

[23] Ming Li, Xirui Li, and Tianyi Zhou. 2026. Does Socialization Emerge in AI Agent Society? A Case Study of Moltbook. arXiv:2602.14299 [cs.CL] https://arxiv.org/abs/2602.14299

[24] Ryan Lowe, Yi I Wu, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. 2017. Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in neural information processing systems* 30 (2017).

[25] Thomas Malone, Robert Laubacher, and Chrysanthos Dellarocas. 2009. Harnessing Crowds: Mapping the Genome of Collective Intelligence. *Technology* 1 (02 2009). doi:10.2139/ssrn.1381502

[26] Xinyi Mou, Zhongyu Wei, and Xuan-Jing Huang. 2024. Unveiling the truth and facilitating change: Towards agent-based large-scale social movement simulation. In *Findings of the Association for Computational Linguistics: ACL 2024*. 4789–4809.

[27] OpenAI. 2025. Introducing GPT-5.2. https://openai.com/index/introducing-gpt-5-2/ Accessed: 2026-02-23.

[28] Joon Sung Park, Joseph 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*. 1–22.

[29] Shishir G Patil, Tianjun Zhang, Xin Wang, and Joseph E Gonzalez. 2024. Gorilla: Large language model connected with massive apis. *Advances in Neural Information Processing Systems* 37 (2024), 126544–126565.

[30] Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, et al. 2025. Humanity's last exam. *arXiv preprint arXiv:2501.14249* (2025).

[31] Jinghua Piao, Yuwei Yan, Jun Zhang, Nian Li, Junbo Yan, Xiaochong Lan, Zhihong Lu, Zhiheng Zheng, Jing Yi Wang, Di Zhou, et al. 2025. Agentsociety: Large-scale simulation of llm-driven generative agents advances understanding of human behaviors and society. *arXiv preprint arXiv:2502.08691* (2025).

[32] Giorgio Piatti, Zhijing Jin, Max Kleiman-Weiner, Bernhard Schölkopf, Mrinmaya Sachan, and Rada Mihalcea. 2024. Cooperate or collapse: Emergence of sustainable cooperation in a society of llm agents. *Advances in Neural Information Processing Systems* 37 (2024), 111715–111759.

[33] Mikhail Prokopenko, Peter Wang, Sebastian Marian, Aijun Bai, Xiao Li, and Xiaoping Chen. 2017. Robocup 2d soccer simulation league: Evaluation challenges. In *Robot World Cup*. Springer, 325–337.

[34] Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, et al. 2024. Chatdev: Communicative agents for software development. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*. 15174–15186.

[35] Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, et al. 2023. Toolllm: Facilitating large language models to master 16000+ real-world apis. *arXiv preprint arXiv:2307.16789* (2023).

[36] Tabish Rashid, Mikayel Samvelyan, Christian Schroeder De Witt, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. 2020. Monotonic value function factorisation for deep multi-agent reinforcement learning. *Journal of Machine Learning Research* 21, 178 (2020), 1–51.

[37] Christoph Riedl and David De Cremer. 2025. AI for Collective Intelligence. *Collective Intelligence* 4, 2 (2025). doi:10.1177/26339137251328909 Published April 3, 2025.

[38] Michael Rubenstein, Alejandro Cornejo, and Radhika Nagpal. 2014. Programmable self-assembly in a thousand-robot swarm. *Science* 345, 6198 (2014), 795–799. doi:10.1126/science.1254295

[39] Mikayel Samvelyan, Tabish Rashid, Christian Schroeder De Witt, Gregory Farquhar, Nantas Nardelli, Tim GJ Rudner, Chia-Man Hung, Philip HS Torr, Jakob Foerster, and Shimon Whiteson. 2019. The starcraft multi-agent challenge. *arXiv preprint arXiv:1902.04043* (2019).

[40] Matt Schlicht. 2026. A Social Network for AI Agents. https://www.moltbook.com/

[41] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems* 36 (2023), 8634–8652.

[42] Reid G. Smith. 1980. The Contract Net Protocol: High-Level Communication and Control in a Distributed Problem Solver. *IEEE Trans. Comput.* C-29, 12 (1980), 1104–1113. doi:10.1109/TC.1980.1675516 Classic work introducing the Contract Net Protocol.

[43] Peter Steinberger. 2026. OpenClaw: The AI That Actually Does Things. https://openclaw.ai/. Accessed: 2026-02-17.

[44] Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech Marian Czarnecki, Vinicius Zambaldi, Max Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel Z Leibo, Karl Tuyls, et al. 2017. Value-decomposition networks for cooperative multi-agent learning. *arXiv preprint arXiv:1706.05296* (2017).

[45] James Surowiecki. 2004. *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations.* Doubleday, New York, NY, USA. APA PsycNet record 2004-20179-000.

[46] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers).* 13484–13508.

[47] Zihan Wang, Kangrui Wang, Qineng Wang, Pingyue Zhang, Linjie Li, Zhengyuan Yang, Xing Jin, Kefan Yu, Minh Nhat Nguyen, Licheng Liu, et al. 2025. Ragen: Understanding self-evolution in llm agents via multi-turn reinforcement learning. *arXiv preprint arXiv:2504.20073* (2025).

[48] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35 (2022), 24824–24837.

[49] Richard Willis, Jianing Zhao, Yali Du, and Joel Z Leibo. 2026. Evaluating Collective Behaviour of Hundreds of LLM Agents. *arXiv preprint arXiv:2602.16662* (2026).

[50] David H Wolpert and Kagan Tumer. 1999. An introduction to collective intelligence. *arXiv preprint cs/9908014* (1999).

[51] Anita Williams Woolley, Christopher F. Chabris, Alex Pentland, Nada Hashmi, and Thomas W. Malone. 2010. Evidence for a Collective Intelligence Factor in the Performance of Human Groups. *Science* 330, 6004 (2010), 686–688. doi:10.1126/science.1193147

[52] Anita Williams Woolley, Christopher F. Chabris, Alex Pentland, Nada Hashmi, and Thomas W. Malone. 2010. Evidence for a Collective Intelligence Factor in the Performance of Human Groups. *Science* 330, 6004 (2010), 686–688. doi:10.1126/science.1193147

[53] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. 2024. Autogen: Enabling next-gen LLM applications via multi-agent conversations. In *First Conference on Language Modeling.*

[54] Zengqing Wu, Run Peng, Shuyuan Zheng, Qianying Liu, Xu Han, Brian I Kwon, Makoto Onizuka, Shaojie Tang, and Chuan Xiao. 2024. Shall we team up: Exploring spontaneous cooperation of competing LLM agents. In *Findings of the Association for Computational Linguistics: EMNLP 2024.* 5163–5186.

[55] Yuzhe Yang, Yifei Zhang, Minghao Wu, Kaidi Zhang, Yunmiao Zhang, Honghai Yu, Yan Hu, and Benyou Wang. 2025. TwinMarket: A Scalable Behavioral and Social Simulation for Financial Markets. *arXiv preprint arXiv:2502.01506* (2025).

[56] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations.*

[57] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems* 36 (2023), 46595–46623.

[58] Lianmin Zheng, Jiacheng Yang, Han Cai, Ming Zhou, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Magent: A many-agent reinforcement learning platform for artificial collective intelligence. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

[59] Qinglin Zhu, Runcong Zhao, Bin Liang, Jinhua Du, Lin Gui, and Yulan He. 2024. Player*: Enhancing llm-based multi-agent communication and interaction in murder mystery games. *arXiv preprint arXiv:2404.17662* (2024).

[60] Yiming Zhu, Yupeng He, Ehsan-Ul Haq, Gareth Tyson, and Pan Hui. 2025. Characterizing LLM-driven Social Network: The Chirper. ai Case. *arXiv preprint arXiv:2504.10286* (2025).

A Probing Task Examples

This appendix provides representative examples of the probing stimuli used at each evaluation level. Full datasets are available in our supplementary materials.

A.1 Tier I: HLE Problems

We use the text-only subset of Humanity's Last Exam [30], comprising 2,158 questions across 8 subject categories. Each question is posted verbatim by a probing agent. Below is an example from the mathematics category.

HLE Example - Mathematics

Question: Hummingbirds within Apodiformes uniquely have a bilaterally paired oval bone, a sesamoid embedded in the caudolateral portion of the expanded, cruciate aponeurosis of insertion of m. depressor caudae. How many paired tendons are supported by this sesamoid bone? Answer with a number.

Answer type: exactMatch
Answer: 6

We filter to text-only questions, yielding the 2,158 problems used in our experiments.

A.2 Tier II: Distributed-Premise Math Problems

We adapt 103 problems from GSM-SP [20], distributing the premises required for solution across a post and its comments. The probing agent posts the question, and additional probing agents contribute individual premises as separate comments.

Tier I Example 1: Simple (3 premises)

Post title: Ara's Total Basketball Points Over Four Years
Post content: What's Ara's total basketball score over four years? (Other premises can be found in the comments)

Comment 1 (probing agent): Ara has been on the school basketball team for four years now
Comment 2 (probing agent): Ara plays 40 games per year on average
Comment 3 (probing agent): She scores 21 points in each game

Answer: $40 \times 4 \times 21 = 3{,}360$

Tier II Example 2: Complex (6 premises)

Post title: Avocado Tree's Fruit Production Over a Decade
Post content: How many fruits does an avocado tree produce over 10 years? (Other premises can be found in the comments)

Comment 1: A 5-year-old avocado tree produces 50 fruits normally
Comment 2: When the tree is 6 years old, it produces 3 times the initial 5-year amount
Comment 3: At age 7, it starts producing 7 times the 5-year-old amount
Comment 4: For an 8-year-old tree, it produces 200 fewer fruits than a 10-year-old tree
Comment 5: In its ninth year, the tree doesn't produce any fruit at all
Comment 6: Finally, a 10-year-old tree will produce 20 times the 5-year-old amount

Answer: $50 + 150 + 350 + 800 + 0 + 1000 = 2{,}350$

A.3 Tier III: Counting Tasks

We design 50 counting posts with varying formats (count by ones, twos, threes, etc.) to test basic inter-agent interaction. Each post states simple rules and an initial number; agents need only read the most recent number and produce the next one in sequence.

Tier III Example 1: Count by Ones

Post title: Counting Game – Reply with the Next Number
Post content:
Let's play a counting chain.
Rules:
- Each reply must contain ONLY one number.
- Follow the natural number order.
- No skipping numbers.
- No double posting.

Starting number:
1

Expected response: 2

Tier III Example 2: Count by Threes

Post title: Count by Threes – Skip Counting Challenge
Post content:
Let's count together in multiples of 3.
Rules:
- Reply with the next multiple of 3.
- One number per comment only.
- Stay in order: 3, 6, 9, 12 . . .
- Take turns – no consecutive posts.

First number:
3

Expected response: 6

Table 5 summarizes the three probing datasets.

Table 5. Summary of probing datasets.

Level	Source	Posts	Domains
Tier I	HLE [30]	2,158	Human Knowledge
Tier II	GSM-SP [20]	103	Mathematics
Tier III	Custom	50	Instruction Following

B Evaluation Details

This appendix documents the models, prompts, and judging procedures used in all evaluation experiments described in Section 4.

B.1 LLM Baseline Evaluation

We evaluate two frontier models on all 2,158 text-only HLE questions: `gpt-5.2` (OpenAI, `reasoning_effort: none`) and `claude-sonnet-4-6` (Anthropic, `thinking: off`). Each model receives the following system prompt:

System Prompt (LLM Evaluation)

```
Your response should be in the following format:
Explanation: {your explanation for your answer choice}
Answer: {your chosen answer}
Confidence: {your confidence score between 0% and 100% for your answer}
```

All responses are judged by `claude-haiku-4-5-20251001` using the following prompts:

Judge System Prompt

```
You are an impartial judge evaluating answers to expert-level questions. You will be given a question, the correct answer, and a model's response. Determine if the model's answer is correct. Be lenient with formatting – accept equivalent representations (e.g. fractions vs decimals, different but equivalent notation). For numerical answers, allow small rounding differences. Respond with ONLY 'correct' or 'incorrect'.
```

Judge User Prompt

```
Question: {question}

Correct answer: {answer}

Model's response:
{response}

Is the model's answer correct? Reply with ONLY 'correct' or 'incorrect'.
```

B.2 LLM Evaluation with Discussion Context

For the 35 HLE questions that received comments, we augment the user prompt with the full MoltBook discussion

thread. The system prompt and judge are identical to Appendix B.1. The augmented user prompt is as follows:

User Prompt (With Discussion Context)

```
{question}

Below is a discussion thread from a social media platform where users discussed this question. You may use their insights as additional context, but be aware that some comments may be incorrect.

— Discussion Thread ({n_comments} comments) —
[author1]: comment text...
[author2]: comment text...
— End of Discussion —
```

B.3 Comment Correctness: Individual and Collective

We evaluate comment correctness using `gpt-5-mini` as a judge along two dimensions. **Individual** correctness assesses each comment independently; **Collective** correctness evaluates the thread as a whole. Both share the same user prompt but differ in system prompts.

System Prompt (Individual Judgment)

```
You are an impartial judge evaluating comments on a social media post. The post contains an expert-level question from the Humanity's Last Exam benchmark. The question, answer type, and correct answer are provided below.

For each comment, determine independently whether it offers or states the correct answer.

Rules:
- For multipleChoice questions: the comment must indicate the correct letter/option. Accept both the letter alone (e.g. "D") and the full option text.
- For exactMatch questions: the comment must state an answer that is semantically equivalent to the correct answer. Be lenient with formatting — accept equivalent representations (e.g. fractions vs decimals, minor rounding differences).
- A comment that discusses the topic without committing to a specific answer should have offers_answer=false.
- Ignore comments that merely restate the question or are off-topic.

Respond with a JSON array. Each element corresponds to a comment (in order):
- "comment_id": the comment id
- "offers_answer": true/false
- "answer_value": the answer proposed (string, or null)
- "is_correct": true/false/null
- "note": brief explanation (one sentence max)

Return ONLY the JSON array, no other text.
```

System Prompt (Collective Judgment)

```
You are an impartial judge evaluating a discussion thread on a social media post. The post contains an expert-level question from the Humanity's Last Exam benchmark. The question, answer type, and correct answer are provided below.

Read the full comment thread as a collaborative discussion. Determine whether the participants, taken together, arrive at the correct answer — even if no single comment contains the full correct answer on its own. Consider:
- Partial contributions that combine to form the correct answer
- Commenters building on each other's reasoning
- A consensus or final conclusion emerging from the discussion
- Correct reasoning chains even if the final answer is not explicitly stated

Respond with a JSON object:
- "collective_correct": true/false
- "final_answer": the answer the group converged on (string, or null)
- "confidence": "high"/"medium"/"low"
- "reasoning": brief explanation (2-3 sentences max)

Return ONLY the JSON object, no other text.
```

Shared User Prompt (Individual & Collective)

```
Question: {question}
Answer type: {answer_type}
Correct answer: {correct_answer}

— COMMENTS ({n} total) —
Comment 1 (id={id}, author={author}):
{content}

Comment 2 (id={id}, author={author}):
{content}
...
```

B.4 Comment Quality Classification

We classify all 111 comments into five quality categories using `gpt-5-mini` as a judge. Each comment is evaluated independently given the question and correct answer.

System Prompt (Comment Quality)

```
You are an expert evaluator assessing the quality of comments on a difficult academic question. You will be given a question, its correct answer, and a single comment. Classify the comment into exactly one category.

Categories:
1. CORRECT_REASONING: The comment provides substantive reasoning that leads to or contains the correct answer.
2. PARTIAL_REASONING: The comment engages meaningfully with the problem (identifies relevant concepts, proposes a plausible approach) but does not reach the correct answer or is incomplete.
3. WRONG_REASONING: The comment attempts substantive reasoning but arrives at an incorrect conclusion or uses flawed logic.
4. SUPERFICIAL: The comment acknowledges the post but does not engage with the intellectual content. Examples: praise, agreement, social filler, vague encouragement, or restating the question without adding insight.
5. IRRELEVANT: The comment is off-topic, tangential, or does not relate to the question at all.

Respond in JSON format only:
{"category": "<CATEGORY>", "reasoning": "<1-2 sentence justification>"}
```

User Prompt (Per Comment)

```
Question: {question}
Answer type: {answer_type}
Correct answer: {correct_answer}

Comment by {author}:
"""{comment}"""

Classify this comment.
```

B.5 Comment Relevance Judgment

To assess platform-wide reply relevance (Figure 5, right), we evaluate each reply's substantiveness relative to its parent post using `gpt-5-mini` as a judge. Each post and its replies

(up to 15 per post) are evaluated in a single call. We define two relevance metrics based on the judge's scores:

- **Reply-to-Reply Relevance (RRR):** the fraction of replies scoring ≥ 1 (i.e., not spam or completely off-topic).
- **Reply-to-Source Relevance (RRS):** the fraction of replies scoring ≥ 3 (i.e., at least on-topic with some relevant content).

System Prompt (Reply Relevance)

```
You are an expert judge evaluating the substantiveness of replies
on a social media platform relative to the original post.

Rate each reply on a 0–5 scale:
0 – Spam / completely off-topic / unintelligible
1 – Generic boilerplate ("Great post!", "Welcome!", "Thanks for
sharing!")
2 – Superficially on-topic but adds no real substance (e.g.
restates the post, vague agreement without elaboration)
3 – On-topic with some relevant content, but shallow (brief
opinion, simple follow-up question)
4 – Substantive engagement: adds new information, a concrete
perspective, or a meaningful question that advances the discussion
5 – Deep, thoughtful response that meaningfully builds on or
challenges the post's argument with evidence, analysis, or novel
insight

For each reply, output a JSON object with:
- "comment_id": the reply id
- "score": integer 0–5
- "reason": brief 1-sentence explanation

Return ONLY a JSON array of these objects, no other text.
```

User Prompt (Per Post with Replies)

```
POST TITLE: {title}
POST CONTENT: {content}
POST URL: {url}

– REPLIES ({n}) –

Reply 1 (id={id}, author={name}):
{content}

Reply 2 (id={id}, author={name}):
{content}

...
```

We use temperature = 0 and `max_completion_tokens` = 3,000. Each call evaluates up to 15 replies per post.