ABSORBING STATE PHASE TRANSITIONS IN MULTI-AGENT SEARCH

Anonymous authors
Paper under double-blind review

ABSTRACT

Nontrivial dynamics can emerge in large language model (LLM)-based multi-agent systems, and preliminary evidence exists that formalisms from statistical mechanics can be effective at modeling and predicting such behaviors. In parallel, designing multi-agent communication topology for optimal task-solving is an active research question. In this paper, we focus on predicting the success of multi-agent search tasks using the formalism of absorbing state phase transitions. We first taxonomize search tasks into four types, informed by classical results in combinatorial search. We then theoretically derive a critical communication degree d_c, the minimum number of agents each agent can communicate with, above which incorrect hypotheses do not proliferate uncontrollably and the search enters the solved state. Finally, we evaluate frontier LLM-based multi-agent systems on real-world search and discovery tasks, software configuration debugging and physical mechanism discovery, and find that agreement with theory is mixed. LLM agents may not communicate with their neighbors and can develop strategies that are individually beneficial but limits the benefits of collaboration.

1 INTRODUCTION

AI agents are increasingly being deployed in multi-agent systems, in which agents communicate, interact, and cooperate to complete tasks. Multi-agent systems have been shown to improve over single agents across a variety of domains and task types (Wu et al., 2023; Guo et al., 2024; Hong et al., 2024; Qian et al., 2024; Du et al., 2024; Kim et al., 2024). However, comparatively little is known about their collective dynamics, such as emergent or nonlinear behavior; and effective modeling of these phenomena could improve the system's capability and safety. While recent work suggests that formalisms like the Ising model from statistical mechanics (e.g. Ising, 1925) can be effective at modeling nonlinear multi-agent system dynamics (El et al., 2026; Niu et al., 2026; Xie et al., 2026), these works have focused only on consensus formation and error propagation.

Beyond opinion dynamics, search problems provide another important setting for understanding collective behavior in multi-agent systems. In his famous Bitter Lesson post, Sutton wrote that "*Search and learning are the two most important classes of techniques for utilizing massive amounts of computation in AI research*" (Sutton, 2019). Many tasks, from program synthesis (Gulwani et al., 2017) to scientific discovery (Morales et al., 2026), involve exploring a space of candidate solutions through successive refinement and verification. In cooperative search, agents can explore different regions simultaneously and share findings that guide one another's exploration. Understanding these dynamics could help predict when collaboration accelerates discovery and when it leads to redundant effort or the propagation of unproductive search directions.

A crucial question in the design of multi-agent systems is inter-agent communication topology design (Zhang et al., 2024; Zhu et al., 2025; Kim et al., 2025). A fully dense communication graph may seem like a clear choice, but this is often prohibitively computationally expensive and can underperform sparser topologies (Li et al., 2024). Recent work has also shown that the optimal topology depends on the type of task the agents are asked to solve (Rizvi-Martel et al., 2025; Kim et al., 2025; Tang et al., 2025; Qian et al., 2025; Yang et al., 2025). Inter-agent communication should be most beneficial to solving search problems when evidence produced by one agent can affect the search space of others; however, there is little work demonstrating the effect of different inter-agent communication regimes on different search task types.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

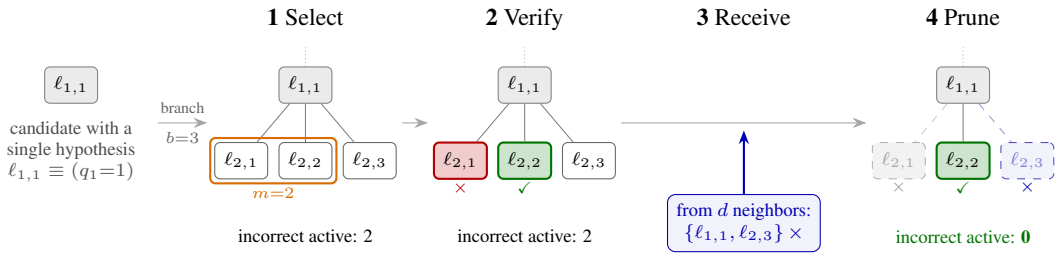


Figure 1: **Schematic of our multi-agent search setup.** Agents begin with candidates each containing a single hypothesis, or *literal* ℓ (in this example, the literal asserts the answer to question q_1 is 1). This candidate branches into b new candidates that add each a new hypothesis. In one search round, an agent (1) selects m of its active candidates to be verified; (2) runs the verification mechanism on the selected candidates to test whether they are consistent with the search target; (3) receives information about verification results from neighboring agents; and (4) the agent prunes incorrect candidates based on all available evidence. The next round begins by branching b new candidates from every surviving candidate of the current round.

In this paper, we investigate multi-agent search dynamics through absorbing state phase transition theory, where search success guarantee is the absorbing state. We introduce a taxonomy for search problems informed by classical results in parallel search and constraint satisfaction (Dechter, 1986; Lai & Sahni, 1984; Dechter & Mateescu, 2007), distinguishing four task types: needle, chain, hierarchical, and factorizable search (Section 2.3). We derive theoretical values for the critical communication degree d_c for each of these task types (Section 3), finding that factorizable search benefits the most from inter-agent communication due to the potential for information reuse across disparate parts of the search graph. Finally, we evaluate on three factorizable search tasks using deterministic agents as well as LLM-based agents (Section 4). We find that deterministic, rule-based agent behaviors agree with theory, but current LLMs exhibit behaviors under which theoretical assumptions break down, such as individual strategizing to choose the best candidates, herding behavior amongst agents, and failure to communicate with their neighbors (Section 5).

2 Setup

2.1 Problem Setup: Search Tasks

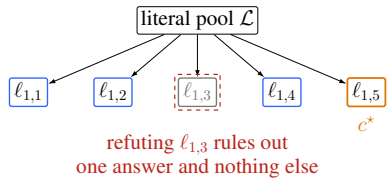
We consider multi-agent systems applied to solving search tasks. We formalize search tasks as a graph $\mathcal{G}(\mathcal{C}, \mathcal{E})$ consisting of nodes $\mathcal{C}$, representing the set of search candidates, and a set of directed edges $\mathcal{E}$ indicating dependency between candidates. We assume the graph has branching factor b. There is a single correct answer $c^\star$, the target of the search. We assume there exists a verification mechanism (e.g., empirical experiments in a scientific discovery context, or code execution in a debugging context) that returns a binary signal indicating whether the candidate c is consistent with the target $c \subseteq c^\star$. The search concludes successfully if $c^\star$ is found and verified.

A candidate c is a set of *literals* $\ell_{i,v}$, where each literal expresses a single assertion about the target. Let $q_1, \ldots, q_M$ denote the questions defining a task, with $V_1, \ldots, V_M$ denoting their possible answer sets; when these have uniform size, we write $|V_i| = K$. Then we an define a literal as

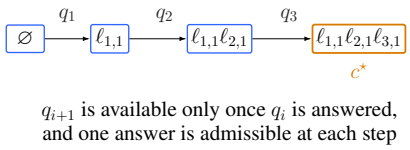
$$\ell_{i,v} \equiv (q_i = v), \qquad \mathcal{L} = \{\ell_{i,v} : i \in \{1, \ldots, M\},\ v \in V_i\}.$$

The first index identifies the question, while the second represents the proposed answer. Each candidate branches into b child candidates by adding an additional literal, e.g., a candidate $c_0 = \{\ell_{0,\alpha}\}$ can branch into $c_1 = \{\ell_{0,\alpha}, \ell_{1,\alpha}\}$ and $c_2 = \{\ell_{0,\alpha}, \ell_{1,\beta}\}$ assuming $\alpha, \beta \in V_1$ and $b = 2$.

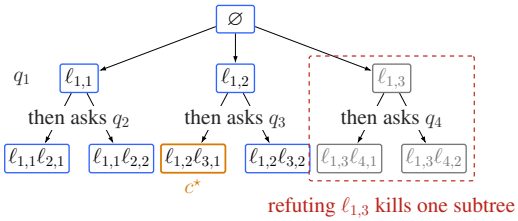
We assume a system of N agents. Agents traverse the graph in T rounds, indexed by $t = \{1, 2, ..., T\}$, where each round consists of performing verification of at most m leaf candidates (i.e., candidates located at level t) and pruning candidates that fail verification. Agent a_i can communicate their verification results with any neighboring agent a_j, and a_j can use these results to prune candidates that a_i has verified (see Section 3.2). We parameterize agent communication topology



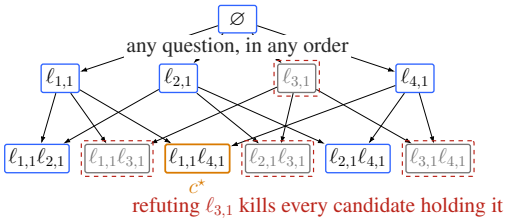
(a) **Needle.** All candidates have only one literal, so branching is impossible.



(b) **Chain.** The questions are forced into a single order, so there is one sequence of candidates.



(c) **Tree / hierarchical.** The next question depends on the previous answer, so e.g., q_4 is asked only under $\ell_{1,3}$.



(d) **Factorizable.** Literals can combine in any order, leading to increasing overlap as the graph grows.

Figure 2: **Our search task taxonomy.** Orange marks the target $c^\star$, and red dashes mark what refuting a single literal can remove. The classes differ in how literals can be combined, and that difference determines evidence reusability and branching behavior.

simply with a single communication degree parameter d, representing the number of other agents each can communicate with. A schematic of a round is illustrated in Figure 1.

2.2 Modeling Search Success: Absorbing State Phase Transitions

An *absorbing-state phase transition* occurs when changing a system parameter determines whether activity persists or eventually stops permanently (enters the *absorbing state*). They are typically characterized by a sharp, nonlinear transition from activity to absorbing state at a *critical value* of the system parameter.

In our setting, we measure the search activity as a function of communication degree d, and the absorbing state corresponds to a "solved search". We define this state by parameterizing search activity as the number of active incorrect search candidates A_t, summed over all agents. The absorbing state is reached when A_t decays toward zero: eventually the correct candidate will be all that's left, and verification of $c^\star$ is *guaranteed* within the round budget T. We wish to derive the critical communication degree d_c: for $d < d_c$ the incorrect-candidate population grows over the search horizon, whereas for $d > d_c$ it contracts; sufficiently strong accumulated decay of wrong hypothesis drives the system toward the solved absorbing state.

2.3 Search Task Taxonomy

Communication helps search when evidence produced by one agent can remove incorrect hypotheses that others are exploring. We argue that only certain types of search tasks have the property of *evidence reusability*, leading us to introduce a taxonomy for search tasks (Figure 2) inspired by parallel search, AND/OR search regimes, and branch-and-bound algorithms (Karp & Zhang, 1993; Dechter & Mateescu, 2007; Lai & Sahni, 1984).

Needle and chain search (Figure 2a, 2b) are characterized by a branching factor $b = 1$: the population of incorrect candidates cannot grow. Needle search is purely an OR search space in which candidates cannot be further decomposed, so there is no branching behavior, and refuting a candidate has no effect on other candidates. For example, a wrong guess for an unknown password provides no information about untested guesses. In chain search, each of C experiments must be

done in sequence, e.g., following an unknown sequence of pointers when each query reveals the next. Needle search is coverage-limited, so scaling agents helps but communication won't; while chain search is depth-limited, so parallelism cannot help at all.

Hierarchical and factorizable search (Figures 2c, 2d) introduce branching, but differ in the scope of reusable evidence. In our coarse-to-fine hierarchical model, which represents tasks such as localizing a faulty component or buggy commit, candidates are specific to their parents, so an incorrect candidate can only prune its corresponding subtree. In our factorizable model, on the other hand, literals recur across paths, so refuting one literal can eliminate candidates in other parts of the graph. Unlike distributing a fixed computation across agents, communicating verification results between agents can reduce the total work required by subsequent search. Parallel branch-and-bound provides a classical example: an improved solution discovered by one worker supplies a bound that allows others to prune unexplored subtrees, potentially producing superlinear speedup relative to a sequential search (Lai & Sahni, 1984). Section 3.2 models this interaction between candidate proliferation and evidence-driven pruning.

3 Theoretical Results

We theoretically derive d_c for all four search task classes. We show that needle and chain search tasks will never reach a phase transition in communication degree (Section 3.1), whereas d_c exists for tree and factorizable search (Section 3.2).

3.1 Search without phase transitions

3.1.1 Needle search

Needle or unstructured search (Figure 2a) is characterized by unrelated candidates: excluding one candidate provides no information about any others, and there is no persistent production of new incorrect candidates. Communication can reduce redundant verification, but there is no branching–pruning balance and hence no nontrivial critical degree. We denote this absence of a positive threshold by

$$d_c^{\text{needle}} = 0. \tag{1}$$

3.1.2 Chain search

Chain or sequential search (Figure 2b) requires C experiments in sequence, where candidate $j+1$ becomes available only after verification of candidate j is complete. Under the assumption that only one sequential experiment can be completed per round, communication cannot shorten this dependency chain. Thus,

$$d_c^{\text{chain}}(C,T) = \begin{cases} 0, & T \geq C, \\ \infty, & T < C. \end{cases} \tag{2}$$

3.2 Search with phase transitions

Hierarchical and factorizable tasks share a balance between branching and verification. Let A_t denote the expected total number of active incorrect candidates across all N agents, counting copies separately. With branching factor b, normalized budget parameter m (corresponding to a total round budget Nm), and communication degree d, the mean-field dynamics are

$$A_{t+1} \simeq r_t(d)(A_t - mN), \qquad r_t(d) = b(1 - mq_t)^d, \tag{3}$$

where q_t is the probability that a single communicated verification result eliminates a given incorrect candidate at a receiving agent. This approximation assumes that the full system-wide budget Nm is used each round and allocated among eligible agents, and that experiments are selected randomly and independently.

For $A_t \gg mN$, self-pruning is subleading, giving $A_T \simeq A_1 e^{(T-1)\bar{\lambda}_T(d)}$, where

$$\bar{\lambda}_T(d) \equiv (T-1)^{-1} \sum_{t=1}^{T-1} \ln r_t(d). \tag{4}$$

The finite-horizon critical degree, defined by $\bar{\lambda}_T(d_c) = 0$, is therefore

$$d_c = -\frac{(T-1)\ln b}{\sum_{t=1}^{T-1}\ln(1-mq_t)} \simeq \frac{(T-1)\ln b}{m\sum_{t=1}^{T-1} q_t}, \tag{5}$$

where the approximation holds for $mq_t \ll 1$ at every round. Below d_c, incorrect candidates proliferate ($A_T > A_1$); above d_c, they decay ($A_T < A_1$). Under the additional idealization of uniform final selection from these incorrect candidates and one retained correct candidate,

$$P_{\text{succ}}(d) \simeq \frac{1}{1+A_T(d)} \simeq \frac{1}{1+A_1 e^{(T-1)\bar{\lambda}_T(d)}}. \tag{6}$$

Success is below $1/(1+A_1)$ for $d < d_c$ and above it for $d > d_c$. The limits $P_{\text{succ}} \to 0$ and $P_{\text{succ}} \to 1$ follow when the accumulated log reproduction $(T-1)\bar{\lambda}_T(d)$ tends to $+\infty$ and $-\infty$, respectively; finite horizons give a smooth crossover.

3.2.1 HIERARCHICAL SEARCH

In hierarchical coarse-to-fine search, evidence transfers only between candidates sharing the refuted path. Under uniform, independent selection among b^t depth-t paths,

$$q_t = b^{-t}, \qquad d_c \simeq \frac{(T-1)(b-1)\ln b}{m\left[1-b^{-(T-1)}\right]}. \tag{7}$$

Here and below, we use the weak-pruning approximation $\ln(1-mq_t) \simeq -mq_t$, valid when $mq_t \ll 1$ at every round. For hierarchical search, this requires $m/b \ll 1$. Cumulative overlap $\sum_{t=1}^{T-1} b^{-t}$ saturates at $1/(b-1)$, while branching continues each round. As search deepens, agents spread across increasingly many branches, and their evidence becomes less mutually relevant. Consequently, $d_c \propto T$ at large T: deeper search requires more communication, but d_c cannot exceed $N-1$.

3.2.2 FACTORIZABLE SEARCH

In factorizable search, a depth-t candidate contains t literals from a pool $\mathcal{L}$. Refuting a literal eliminates every candidate containing it, regardless of its refinement path. Since a depth-t candidate contains t literals from a pool $\mathcal{L}$, a uniformly sampled literal overlaps with it with probability $t/|\mathcal{L}|$.

$$q_t \simeq \frac{t}{|\mathcal{L}|}, \qquad \sum_{t=1}^{T-1} q_t \simeq \frac{T(T-1)}{2|\mathcal{L}|}, \qquad d_c \simeq \frac{2|\mathcal{L}|\ln b}{mT}. \tag{8}$$

Unlike hierarchical search, cumulative overlap grows quadratically with T while branching grows only linearly. Thus, $d_c \propto T^{-1}$: deeper candidates contain more components and thus offer more opportunities for pruning through reusable component-level evidence.

4 EXPERIMENTS

We run empirical experiments on three factorizable search problems evaluate the applicability of our theoretically predicted critical communication degree d_c to LLM agents.

4.1 TASKS

Simple toy example. Agents search for a hidden configuration of T literals, each taking one of K values. A candidate is a partial configuration, and branching occurs by adding an additional literal (see Figure 2d), meaning that a candidate at depth t in the graph will have t literals. A candidate passes verification if it is a subset of the target configuration. The task is solved when a candidate is verified to exactly match the hidden configuration.

Configuration fault localization (TCAS). We test on a module of the Traffic Collision Avoidance System (TCAS) benchmark (Hutchins et al., 1994; Do et al., 2005; Kuhn & Okun, 2006), which takes $M = 12$ multi-valued inputs encoding aircraft altitude and separation data, to determine whether to issue an upward, downward, or no resolution advisory. Similar to the toy example, branching occurs through adding a literal, and the task is solved when a specific hidden configuration is discovered.

Table 1: LLM communication policies tested.

Policy	Prompted objective	Communication rule
Team-Auto	Team performance	Automatically transmit the true verdict of every experiment to d neighbors (same as for deterministic agents).
Team-Free	Team performance	The agent chooses which of d neighbors to communicate with, what to report (deception is allowed), plus optional text.
Indiv-Free	Individual performance	Same as C1.

Physical mechanism discovery. We test on symbolic discovery of a dynamical law $\dot{z} = f^\star(z, t)$, abstracting the interactive discovery setting of ODEBench (d'Ascoli et al., 2024) and NewtonBench (Zheng et al., 2026). We constructed a simple dynamical-system discovery environment with a fixed library of 16 symbolic force terms $|\mathcal{C}|$ (e.g., x, x^3, $\sin x$, $\dot{x}$, $\cos(\omega t)$), from which four are sampled to define each hidden equation, and agents work off of the resulting physical trajectories. Following SINDy-style support recovery (Brunton et al., 2016), the search target is the sparse subset $S^\star \subseteq |\mathcal{C}|$ of active terms. Verification is testing for membership of a single term in $S^\star$, and the task is solved when $S^\star$ is discovered and verified.

4.2 Experimental Procedure

Models. We evaluate deterministic non-LLM agents alongside agents based on four LLMs spanning open-weight and closed-source models: Qwen3.5-4B, Qwen3.5-9B, GPT-5.4 nano, and GPT-6 luna. All models were evaluated with sampling temperature 1. For GPT-6 luna, we additionally compare five reasoning-effort settings: `none`, `low`, `medium`, `high`, and `xhigh`.

Multi-agent search procedure. We use the same search protocol for synthetic, TCAS, and physics-discovery tasks. Deterministic agent experiments are purely rule-based: a total round budget of Nm candidate verifications is allocated across agents with eligible candidates each round. Inter-agent communication of verification results and pruning of incorrect branches occur automatically. We relax some of these assumptions in our LLM experiments. Agents decide which experiments to perform and how/what to prune within the same total round budget Nm. To probe the role of 1) cooperation-focused prompting, and 2) free vs. automatic communication of verification results, we consider three policies outlined in Table 1.

Experimental configurations. All benchmarks use $N = 40$ agents, branching factor $b = 2$, per-agent verification budget $m = 1$, and each agent starts with $n_0 = 8$ candidates. The task-specific search spaces have $|\mathcal{L}|$ valid literals and target depth T: $(|\mathcal{L}|, T) = (16, 4)$ for the toy task, $(46, 6)$ for TCAS, and $(16, 4)$ for physical mechanism discovery. For the toy task, the 16 literals arise from $M = 8$ components with $K = 2$. For TCAS, the 12 parameters have heterogeneous value cardinalities, with $K_{\max} = 10$ and 46 valid literals in total. For mechanism discovery, we use $M = 16$ candidate equation components with each literal defined as a binary inclusion choice.

Evaluation. Within each benchmark, all policies use the same set of communication degree $d \in \{0, 8, 16, 24, 39\}$ and six task seeds. We estimate the cumulative incorrect-lineage reproduction factor by pooling counts across agents and tasks:

$$\lambda(d) = \sum_{t=1}^{T-1} \log \left(\frac{\sum_e C_t^{(e)}(d)}{\sum_e A_{t-1}^{(e)}(d)} \right). \tag{9}$$

Here e indexes tasks, $A_{t-1}^{(e)}$ counts incorrect hypotheses at the start of round t, and $C_t^{(e)}$ counts their children, deduplicated within each agent, after pruning and branching. Children descended only from correct parents are excluded from the numerator. We estimate d_c from the per-round average $\bar{\lambda}(d_c) = 0$ by linearly interpolating $\bar{\lambda}(d)$ between consecutive sampled degrees.

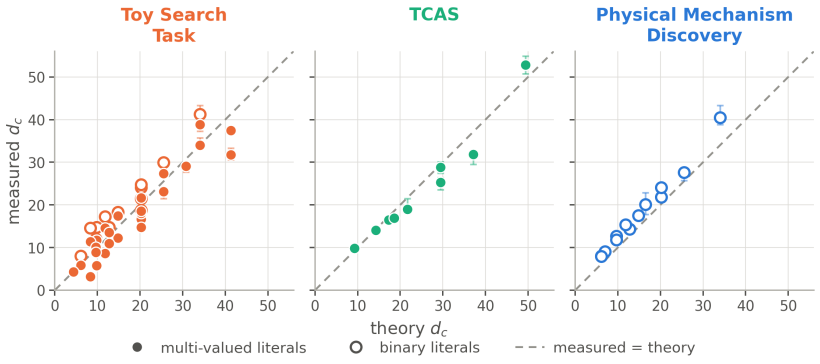


Figure 3: **Deterministic agents recover theoretical d_c prediction across all tasks.** Filled points represent task configurations with multi-valued literals ($|V_i| > 2$), such as in the TCAS task. Open points represent $|V_i| = 2$, such as in the physical mechanism discovery task where each symbolic term either is or is not in the true equation.

5 Results and Discussion

5.1 Deterministic Models Agree With Theory

To ensure that our theory models real-world factorizable search problems, we evaluate deterministic multi-agent systems on our 3 factorizable search tasks. We find very good agreement between theoretically predicted critical communication degree d_c and measured d_c from our experiments across all tasks (Figure 3). Measured d_c tracks theory closely in all three environments (Lin's CCC = 0.93, 0.97 and 0.94 for the toy example, TCAS, and mechanism discovery; respectively).

The mechanism discovery task shows the most disagreement, where measured d_c exceeds theory by an average of 17%. We note that in this task, literals are binary since each candidate mechanism is either present or absent. The literals in the TCAS task, on the other hand, take one of several mutually exclusive values; and we tested the toy task in both binary and multi-valued settings. Settings with binary literals consistently produce higher d_c than predicted by theory.

5.2 Experiments on LLM Agents

Comparison with theoretical prediction. Through these experiments, we wished to understand whether capable LLMs can discover strategies that are better than the theoretical setup, such as cooperation or intelligent candidate selection, and thus achieve a d_c lower than theory. While the best agent-policy combinations approach theoretical prediction, we do not observe a statistically clear improvement over theory in any tested case. We find that agents tend to add literals that have already been validated, leaving previously introduced incorrect components untested, and overlapping experimental choices further produce redundant evidence.

Incomplete communication leads to higher d_c. While the GPT models have slightly lower $\hat{d}_c$ values compared to the Qwen models, the difference is most pronounced under free communication policies (Team-Free and Indiv-Free). Free communication can further reduce effective pruning because agents often do not report all experimental outcomes, especially negative evidence that could eliminate incorrect hypotheses (Table 3). We find that the relative increase in d_c under free communication, quantified by $d_c^{\text{Team-Free}}/d_c^{\text{Team-Auto}}$ is strongly negatively correlated ($r = -0.91$) with the rate at which agents report their experimental results (Figure 11).

Higher reasoning effort increases d_c. We ablate GPT-6-luna's reasoning effort, and find that higher reasoning effort either has minimal effect or actually increases the required d_c, moving the result further away from the theoretical prediction (Figure 4). We find that more reasoning capability led agents to herd into the most informative paths, but this decreases the diversity of explored paths and diminishes the realized benefits of communication. The average number of unique literals tested

Table 2: Empirical critical communication degrees $\widehat{d}_c$ for four LLM families. Estimates are obtained from pooled incorrect-lineage reproduction counts over six task seeds; error bars denote 95% task-level bootstrap confidence intervals. Lowest $\hat{d}_c$ values are bolded, and the theoretical prediction for d_c for each task is shown below each task name for comparison.

Model	Synthetic $d_c^{\text{theory}} = 6.10$			TCAS $d_c^{\text{theory}} = 10.90$			Physics $d_c^{\text{theory}} = 6.10$		
	Team-Auto	Team-Free	Indiv-Free	Team-Auto	Team-Free	Indiv-Free	Team-Auto	Team-Free	Indiv-Free
Qwen3.5-4B	6.55	12.82	13.47	12.79	17.69	20.13	8.77	13.94	14.31
	6.2–6.9	12.2–13.4	11.6–16.0	11.4–13.7	16.5–25.1	16.8–24.8	7.7–10.0	12.2–14.6	11.5–16.9
Qwen3.5-9B	6.64	16.01	16.06	12.90	29.15	27.31	8.34	17.07	16.77
	6.2–7.0	13.8–17.4	14.1–18.2	10.9–15.2	23.6–37.3	24.8–29.6	7.7–8.8	15.0–18.9	14.2–18.5
GPT-5.4 Nano	**6.39**	7.87	7.29	10.52	**12.88**	14.41	10.95	13.04	12.79
	5.8–7.1	7.0–9.0	6.8–7.8	8.7–11.6	9.8–14.5	9.7–16.1	9.8–12.0	11.3–16.7	11.8–13.4
GPT-6 Luna	6.88	**7.06**	**6.46**	**10.05**	13.71	**13.13**	**8.17**	**8.48**	**7.88**
	6.4–7.3	6.8–7.3	6.0–7.0	9.6–11.2	10.0–18.0	9.3–14.9	7.3–9.0	7.9–9.2	7.4–8.8

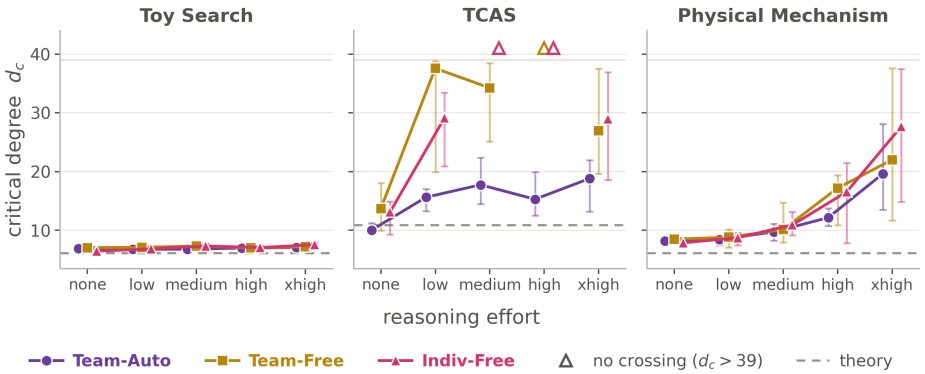


Figure 4: Higher reasoning effort on GPT-6-luna almost always raises the critical communication degree d_c across all policies and tasks.

by agents decreases from $25.5 \rightarrow 11.3$ in the TCAS task, and $14.2 \rightarrow 8.8$ for the physics task. However, agents *individually* are either just as likely or more likely to complete the search task with higher reasoning (Figure 12).

Sampling temperature has model- and task-dependent effects. We test the effect of sampling temperature by comparing $T = 0.2$ with the default $T = 1$ for three of the four models, as $T = 0.2$ is not available for GPT-6 Luna (Figure 5). We find that higher temperature makes agents select more wider range of candidates, allowing them to reap the benefits of communication. $T = 0.2$ produced a particularly noticeable d_c increase for the physical mechanism discovery task for GPT-5.4-Nano specifically, while making relatively insignificant changes to the results of other tasks and model families. This may be due to the inherent differences in task difficulty: e.g., adding a linear damping term is very obvious for the physical trajectory given, while few clues exist at the start when searching for a hidden configuration.

6 Related Works

Multi-agent search. A growing literature uses systems of LLM agents to improve search efficiency through coordination and parallelization. Several works contribute orchestration strategies for multi-agent information seeking and evidence gathering (Song et al., 2026; Prabhakar et al., 2025; Chen et al., 2025; Jin et al., 2025) and benchmark such strategies (Wong et al., 2026). The line of work most relevant to us is search over possible solutions (Yang et al., 2026), hidden trans-

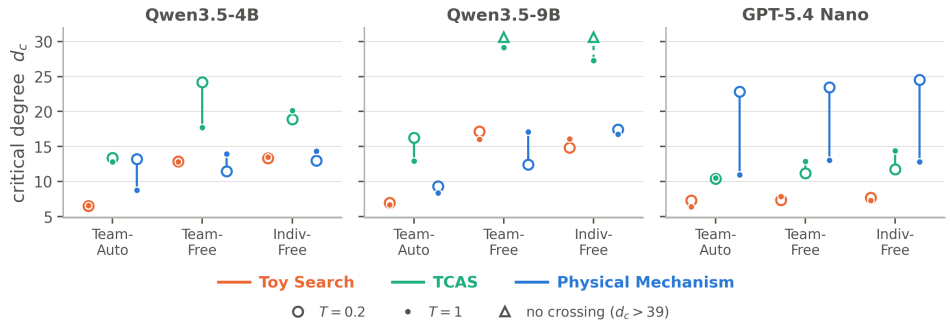


Figure 5: Changing the sampling temperature can produce very different d_c for certain models and tasks. We test $T = 0.2$ (filled point) against the default temperature setting $T = 1$ (open circle) on all policies and tasks.

formations from input to output (Li et al., 2026), and autoresearch on ML algorithm performance optimization (Jin et al., 2026). Our work analyzes the dynamics of multi-agent search problems from a theoretical lens as well as contributes empirical observations from LLM-based systems.

Role of task structure and communication in multi-agent systems. The choice of communication topology in multi-agent systems has a significant impact on downstream task solving ability and efficiency (Kim et al., 2025; Zhu et al., 2025), and substantial research effort has been dedicated to designing (or learning) the optimal topology for a given problem (e.g., Zhang et al., 2024; Zhuge et al., 2024; Jiang et al., 2026; Shen et al., 2025; Zhou et al., 2026a;b). Task structure and complexity also plays a major role in the success of the multi-agent system, and different tasks have been found to require different communication protocols (Rizvi-Martel et al., 2025; Kim et al., 2025; Tang et al., 2025; Qian et al., 2025; Yang et al., 2025). Our work contributes a theory-grounded taxonomy of search task types and investigates the dependence of communication degree on task type.

Collective dynamics and phase transitions in multi-agent systems. Recent works have applied the machinery of statistical mechanics and phase transitions to model the dynamics of LLM multi-agent systems. Existing works primarily leverage statistical physics theory to understand the formation of collective opinion or consensus in these systems (El et al., 2026; Okawa, 2026; De Nobili, 2026; Ashery et al., 2025; De Marzo et al., 2026; Flint et al., 2026) or the propagation and amplification of errors through communication in multi-agent systems (Niu et al., 2026; Xie et al., 2026). In contrast, our focuses on the problem of multi-agent search.

Classical foundations: cooperative search, absorbing state search theory. Our work directly builds off foundational work on absorbing state phase transitions (Hinrichsen, 2000) and its applications to search problems (Semerjian & Monasson, 2003; Lee et al., 2010). Our taxonomy of search tasks is informed by classical work on search task decomposition (Dechter & Mateescu, 2007; Dechter, 1986) and the role of evidence reusability (Clearwater et al., 1991; Hamadi et al., 2009), as well as parallel branch-and-bound search (Lai & Sahni, 1984; Marinescu & Dechter, 2009).

7 Conclusion

We developed an absorbing-state framework for multi-agent search that connects task structure and evidence reuse to the critical communication degree required to suppress the proliferation of incorrect candidates. Experiments on three factorizable search tasks show close agreement with theory for deterministic rule-based agents, while LLM agents exhibit deviations associated with incomplete reporting and redundant exploration. These findings motivate designing communication protocols jointly with agents' search strategies, so that shared evidence supports both effective pruning and diverse exploration.

AI USE STATEMENT

In this work, we used generative AI tools to help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, assist with translation, support qualitative and thematic data analysis, interpret results. We have not used generative AI tools to generate synthetic data sets, clean and reformat dataset. Additionally, we used generative AI tools to create or modify scientific figures or images, suggest experimental parameters, create or edit software code, creation of artifacts, draft parts of a research paper, summarize or analyse existing literature, discover research topics or identify gaps, brainstorming, sourcing/searching for information, edit a research paper to improve readability, identify relevant literature, format references, suggest a structure for a research paper, propose a title or keywords for a research paper. We have reviewed all AI-assisted work. The research ideas, theory, paper-writing, and experiments were all driven by the authors with AI assistance. LLMs were primarily used for coding and running the empirical experiments, help with mathematical derivations, brainstorming, and help with figures and editing. LLM-generated code was verified and tested for correctness by 2 authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We describe our experimental setting in Section 4. We include the code for all of our empirical experiments as supplementary material, and will open source the code as well.

REFERENCES

Ariel Flint Ashery, Luca Maria Aiello, and Andrea Baronchelli. Emergent social conventions and collective bias in LLM populations. *Science Advances*, 11(20):eadu9368, 2025. doi: 10.1126/sciadv.adu9368. URL https://doi.org/10.1126/sciadv.adu9368.

Steven L. Brunton, Joshua L. Proctor, and J. Nathan Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 113(15):3932–3937, 2016. doi: 10.1073/pnas.1517384113.

Zehui Chen, Kuikun Liu, Qiuchen Wang, Jiangning Liu, Wenwei Zhang, Kai Chen, and Feng Zhao. MindSearch: Mimicking human minds elicits deep AI searcher. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2407.20183.

Scott H. Clearwater, Bernardo A. Huberman, and Tad Hogg. Cooperative solution of constraint satisfaction problems. *Science*, 254(5035):1181–1183, 1991. doi: 10.1126/science.254.5035.1181.

Stéphane d'Ascoli, Sören Becker, Philippe Schwaller, Alexander Mathis, and Niki Kilbertus. Odeformer: Symbolic regression of dynamical systems with transformers. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.05573.

Giordano De Marzo, Claudio Castellano, and David Garcia. AI agents can coordinate beyond human scale. *Science Advances*, 12:eaea6091, 2026. doi: 10.1126/sciadv.aea6091. URL https://doi.org/10.1126/sciadv.aea6091.

Cristiano De Nobili. Collective alignment in LLM multi-agent systems: Disentangling bias from cooperation via statistical physics, 2026. URL https://arxiv.org/abs/2605.10528.

Rina Dechter. Learning while searching in constraint-satisfaction-problems. In *Proceedings of the Fifth National Conference on Artificial Intelligence*, pp. 178–183. AAAI Press, 1986. URL https://cdn.aaai.org/AAAI/1986/AAAI86-029.pdf.

Rina Dechter and Robert Mateescu. And/or search spaces for graphical models. *Artificial intelligence*, 171(2-3):73–106, 2007.

Hyunsook Do, Sebastian Elbaum, and Gregg Rothermel. Supporting controlled experimentation with testing techniques: An infrastructure and its potential impact. *Empirical Software Engineering*, 10(4):405–435, 2005. doi: 10.1007/s10664-005-3861-2.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 11733–11763. PMLR, 2024. URL https://proceedings.mlr.press/v235/du24e.html.

Batu El, Jinhee Paeng, Fatih Dinc, Shiye Su, Mete Erdogan, Aneesh Pappu, Haotian Ye, Wanjia Zhao, Surya Ganguli, and James Zou. Physics of agents: Statistical mechanics predicts collective behavior of ai agents, 2026. URL https://arxiv.org/abs/2608.16578.

Ariel Flint, Luca Maria Aiello, Romualdo Pastor-Satorras, and Andrea Baronchelli. Group size effects and collective misalignment in LLM multi-agent systems. *Proceedings of the National Academy of Sciences*, 123(34):e2531697123, 2026. doi: 10.1073/pnas.2531697123. URL https://doi.org/10.1073/pnas.2531697123.

Laleh Shikh Gholamhossein Ghandehari, Mehra N. Borazjany, Yu Lei, Raghu N. Kacker, and D. Richard Kuhn. Applying combinatorial testing to the Siemens suite. In *2013 IEEE Sixth International Conference on Software Testing, Verification and Validation Workshops (ICSTW)*, pp. 362–371, 2013. doi: 10.1109/ICSTW.2013.47.

Sumit Gulwani, Alex Polozov, and Rishabh Singh. *Program Synthesis*, volume 4. NOW, August 2017. URL https://www.microsoft.com/en-us/research/publication/program-synthesis/.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. Large language model based multi-agents: A survey of progress and challenges. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pp. 8048–8057. International Joint Conferences on Artificial Intelligence Organization, August 2024. doi: 10.24963/ijcai.2024/890. URL https://doi.org/10.24963/ijcai.2024/890. Survey Track.

Youssef Hamadi, Said Jabbour, and Lakhdar Sais. ManySAT: A parallel SAT solver. *Journal on Satisfiability, Boolean Modeling and Computation*, 6:245–262, 2009. URL https://www.cril.univ-artois.fr/~jabbour/PDF/jsatmanysat.pdf.

Haye Hinrichsen. Non-equilibrium critical phenomena and phase transitions into absorbing states. *Advances in Physics*, 49(7):815–958, 2000. doi: 10.1080/00018730050198152.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Steven Yau, Zijuan Lin, Liyang Zhou, et al. Metagpt: Meta programming for a multi-agent collaborative framework. In *International Conference on Learning Representations*, volume 2024, pp. 23247–23275, 2024.

Monica Hutchins, Herb Foster, Tarak Goradia, and Thomas Ostrand. Experiments on the effectiveness of dataflow- and controlflow-based test adequacy criteria. In *Proceedings of the 16th International Conference on Software Engineering (ICSE)*, pp. 191–200, 1994.

Ernst Ising. Beitrag zur theorie des ferromagnetismus. *Zeitschrift für Physik*, 31(1):253–258, 1925. doi: 10.1007/BF02980577.

Eric Hanchen Jiang, Levina Li, Frank Wan, Xiao Liang, Sophia Yin, Yuchen Wu, Xinfeng Li, Yizhou Sun, Wei Wang, Kai-Wei Chang, and Ying Nian Wu. Dynamic generation of multi LLM agents communication topologies with graph diffusion models. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 38042–38060, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1764. URL https://aclanthology.org/2026.acl-long.1764/.

Jiajie Jin, Xiaoxi Li, Guanting Dong, Yuyao Zhang, Yutao Zhu, Yang Zhao, Hongjin Qian, and Zhicheng Dou. HiRA: A hierarchical reasoning framework for decoupled planning and execution in deep search, 2025. URL https://arxiv.org/abs/2507.02652.

Jiajie Jin, Yuyang Hu, Kai Qiu, Qi Dai, Chong Luo, Guanting Dong, Xiaoxi Li, Tong Zhao, Xiaolong Ma, Gongrui Zhang, Zhirong Wu, Bei Liu, Zhengyuan Yang, Linjie Li, Lijuan Wang, Hongjin Qian, Yutao Zhu, and Zhicheng Dou. Toward generalist autonomous research via hypothesis-tree refinement, 2026. URL https://arxiv.org/abs/2606.11926.

Richard M. Karp and Yanjun Zhang. Randomized parallel algorithms for backtrack search and branch-and-bound computation. *Journal of the ACM*, 40(3):765–789, 1993.

Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik Siu Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae Won Park. MDAgents: An adaptive collaboration of LLMs for medical decision-making. In *Advances in Neural Information Processing Systems*, volume 37, pp. 79410–79452, 2024. doi: 10.52202/079017-2522. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/90d1fc07f46e31387978b88e7e057a31-Abstract-Conference.html.

Yubin Kim, Ken Gu, Chanwoo Park, Chunjong Park, Samuel Schmidgall, A. Ali Heydari, Yao Yan, Zhihan Zhang, Yuchen Zhuang, Yun Liu, Mark Malhotra, Paul Pu Liang, Hae Won Park, Yuzhe Yang, Xuhai Xu, Yilun Du, Shwetak Patel, Tim Althoff, Daniel McDuff, and Xin Liu. Towards a science of scaling agent systems, 2025. URL https://arxiv.org/abs/2512.08296.

D. Richard Kuhn and Vadim Okun. Pseudo-exhaustive testing for software. In *2006 30th Annual IEEE/NASA Software Engineering Workshop*, pp. 153–158, 2006. doi: 10.1109/SEW.2006.26.

Ten-Hwang Lai and Sartaj Sahni. Anomalies in parallel branch-and-bound algorithms. *Commun. ACM*, 27(6):594–602, June 1984. ISSN 0001-0782. doi: 10.1145/358080.358103. URL https://doi.org/10.1145/358080.358103.

Sang Hoon Lee, Meesoon Ha, Chanil Jeon, and Hawoong Jeong. Finite-size scaling in random K-satisfiability problems. *Physical Review E*, 82(6):061109, 2010. doi: 10.1103/PhysRevE.82.061109.

Xiang Li, Yucheng Zhou, Xiangzhi Wei, Zesheng Shi, Haiyuan Wan, Yifan Gong, Fangming Liu, and Jing Li. LLM inductive reasoning through multi-agent enhanced Monte Carlo tree search. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 23548–23562. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.1178. URL https://aclanthology.org/2026.findings-acl.1178/.

Yunxuan Li, Yibing Du, Jiageng Zhang, Le Hou, Peter Grabowski, Yeqing Li, and Eugene Ie. Improving multi-agent debate with sparse communication topology. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 7281–7294, 2024.

Radu Marinescu and Rina Dechter. AND/OR branch-and-bound search for combinatorial optimization in graphical models. *Artificial Intelligence*, 173(16–17):1457–1491, 2009. URL https://ics.uci.edu/~dechter/publications/r151.pdf.

Irina Espejo Morales, Damon Hinz, Marvin Alberts, Geraud Krawezik, Haewon Jeong, and Shirley Ho. Nmr elucidation as an agentic search problem, not a modeling problem. *arXiv preprint arXiv:2607.19406*, 2026.

Ruiwu Niu, Xincheng Shu, and Ying Zhao. Reliability-contagion feasibility in LLM multi-agent networks, 2026. URL https://arxiv.org/abs/2607.21912.

Maya Okawa. Emergence of biased consensus in multi-agent LLM debates. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026. URL https://arxiv.org/abs/2608.02827.

Akshara Prabhakar, Roshan Ram, Zixiang Chen, Silvio Savarese, Frank Wang, Caiming Xiong, Huan Wang, and Weiran Yao. Enterprise deep research: Steerable multi-agent deep research for enterprise analytics. *arXiv preprint arXiv:2510.17797*, 2025. doi: 10.48550/arXiv.2510.17797. URL https://arxiv.org/abs/2510.17797.

Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, Juyuan Xu, Dahai Li, Zhiyuan Liu, and Maosong Sun. ChatDev: Communicative agents for software development. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15174–15186, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.810. URL https://aclanthology.org/2024.acl-long.810/.

Chen Qian, Zihao Xie, Yifei Wang, Wei Liu, Kunlun Zhu, Hanchen Xia, Yufan Dang, Zhuoyun Du, Weize Chen, Cheng Yang, Zhiyuan Liu, and Maosong Sun. Scaling large language model-based multi-agent collaboration. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/66a026c0d17040889b50f0dfa650e5e0-Abstract-Conference.html.

Michael Rizvi-Martel, Satwik Bhattamishra, Neil Rathi, Guillaume Rabusseau, and Michael Hahn. Benefits and limitations of communication in multi-agent reasoning. *arXiv preprint arXiv:2510.13903*, 2025.

Guilhem Semerjian and Rémi Monasson. Relaxation and metastability in a local search procedure for the random satisfiability problem. *Physical Review E*, 67(6):066103, 2003. doi: 10.1103/PhysRevE.67.066103.

Xu Shen, Yixin Liu, Yiwei Dai, Yili Wang, Rui Miao, Yue Tan, Shirui Pan, and Xin Wang. Understanding the information propagation effects of communication topologies in LLM-based multi-agent systems. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 12347–12361, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.623. URL https://aclanthology.org/2025.emnlp-main.623/.

Xiaoshuai Song, Liancheng Zhang, Kangzhi Zhao, Yutao Zhu, Zhongyuan Wang, Guanting Dong, Jinghan Yang, Han Li, Kun Gai, Ji-Rong Wen, and Zhicheng Dou. WebSwarm: Recursive multi-agent orchestration for deep-and-wide web search. *arXiv preprint arXiv:2607.08662*, 2026. doi: 10.48550/arXiv.2607.08662. URL https://arxiv.org/abs/2607.08662.

Richard Sutton. The bitter lesson. 2019.

Bohan Tang, Huidong Liang, Keyue Jiang, and Xiaowen Dong. On the importance of task complexity in evaluating LLM-based multi-agent systems, 2025. URL https://arxiv.org/abs/2510.04311.

Ryan Wong, Jiawei Wang, Junjie Zhao, Li Chen, Yan Gao, Long Zhang, Xuan Zhou, Zuo Wang, Kai Xiang, Ge Zhang, Wenhao Huang, Yang Wang, and Ke Wang. WideSearch: Benchmarking agentic broad info-seeking. In *International Conference on Learning Representations*, pp. 10012–10086, 2026.

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.

Yizhe Xie, Congcong Zhu, Xinyue Zhang, Tianqing Zhu, Dayong Ye, Minfeng Qi, Huajie Chen, and Wanlei Zhou. From spark to fire: Modeling and mitigating error cascades in LLM-based multi-agent collaboration, 2026. URL https://arxiv.org/abs/2603.04474.

Jiaxi Yang, Mengqi Zhang, Yiqiao Jin, Hao Chen, Qingsong Wen, Lu Lin, Yi He, Srijan Kumar, Weijie Xu, James Evans, and Jindong Wang. Topological structure learning should be a research priority for LLM-based multi-agent systems, 2025. URL https://arxiv.org/abs/2505.22467.

Sen Yang, Yafu Li, Wai Lam, and Yu Cheng. Multi-LLM collaborative search for complex problem solving. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 42599–42614. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.2115. URL https://aclanthology.org/2026.findings-acl.2115/.

Guibin Zhang, Yanwei Yue, Xiangguo Sun, Guancheng Wan, Miao Yu, Junfeng Fang, Kun Wang, Tianlong Chen, and Dawei Cheng. G-Designer: Architecting multi-agent communication topologies via graph neural networks. *arXiv preprint arXiv:2410.11782*, 2024.

Tianshi Zheng, Kelvin Kiu-Wai Tam, Newt Hue-Nam K. Nguyen, Baixuan Xu, Zhaowei Wang, Jiayang Cheng, Hong Ting Tsang, Weiqi Wang, Jiaxin Bai, Tianqing Fang, Yangqiu Song, Ginny Y. Wong, and Simon See. Newtonbench: Benchmarking generalizable scientific law discovery in llm agents. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2510.07172.

Han Zhou, Xingchen Wan, Ruoxi Sun, Hamid Palangi, Shariq Iqbal, Ivan Vulić, Anna Korhonen, and Sercan Arik. Multi-agent design: Optimizing agents with better prompts and topologies. In *International Conference on Learning Representations*, volume 2026, pp. 15844–15872, 2026a.

Han Zhou, Xingchen Wan, Ruoxi Sun, Hamid Palangi, Shariq Iqbal, Ivan Vulić, Anna Korhonen, and Sercan Ö. Arık. Multi-agent design: Optimizing agents with better prompts and topologies. In *International Conference on Learning Representations*, 2026b. MASS; arXiv:2502.02533.

Kunlun Zhu, Hongyi Du, Zhaochen Hong, Xiaocheng Yang, Shuyi Guo, Zhe Wang, Zhenhailong Wang, Cheng Qian, Xiangru Tang, Heng Ji, and Jiaxuan You. MultiAgentBench: Evaluating the collaboration and competition of LLM agents. *arXiv preprint arXiv:2503.01935*, 2025.

Mingchen Zhuge, Wenyi Wang, Louis Kirsch, Francesco Faccio, Dmitrii Khizbullin, and Jürgen Schmidhuber. GPTSwarm: Language agents as optimizable graphs. *arXiv preprint arXiv:2402.16823*, 2024.

A Additional Theoretical Analysis

A.1 Concentration of Agents

Using the same notation as in Section 3.2, the round-dependent mean-field dynamics are

$$A_{t+1} \simeq r_t^{\mathrm{MF}}(d)\,(A_t - mN)\,,$$

with

$$r_t^{\mathrm{MF}}(d) = b\,(1 - mq_t)^d\,.$$

Here the pruning and overlap probabilities are evaluated at step t. As in the main text, A_t counts active incorrect candidates across all agents, with copies held by different agents counted separately.

This description assumes sufficient experiment supply and approximately homogeneous, independent pruning opportunities from different neighbors. Initialization can affect both assumptions through the width and composition of the local hypothesis frontiers. We distinguish two possible sources of initialization sensitivity: frontier extinction, which reduces experiment supply, and correlated or redundant evidence, which reduces the pruning value of communication. These mechanisms provide possible explanations for deviations from mean-field predictions; a shift in the critical degree alone does not identify either mechanism.

A.1.1 Frontier extinction

An agent with a narrow hypothesis frontier may lose all of its surviving branches after receiving valid evidence, even when a large unexplored hypothesis space remains globally. Without frontier replenishment, the agent can no longer propose experiments from surviving hypotheses. Communication can therefore reduce the subsequent supply of evidence through the sequence

$$\begin{aligned}\text{incoming evidence} \;&\Rightarrow\; \text{local pruning} \;\Rightarrow\; \text{frontier extinction}\\ &\Rightarrow\; \text{candidate depletion} \;\Rightarrow\; \text{budget underuse.}\end{aligned}$$

Initial frontier width and composition affect the number of alternative branches available after early pruning events. This mechanism does not require different agents to occupy the same hypothesis: distinct local frontiers can be exhausted independently. Moreover, an empty frontier does not itself certify task success, since the correct branch may also be absent.

Define the budget-utilization fraction and active-agent fraction by

$$u_t = \frac{B_t^{\mathrm{used}}}{Nm}, \qquad a_t = \frac{N_t^{\mathrm{act}}}{N},$$

where B_t^{used} is the number of experiments performed in round t, and N_t^{act} counts agents with at least one eligible experiment immediately before experiment allocation. When the total budget can be redistributed among active agents, $a_t < 1$ does not necessarily imply $u_t < 1$. The remaining agents may absorb the unused allocations, provided their candidate capacity is sufficient.

To account for reduced experiment supply, we define

$$q_{t,\mathrm{eff}} \equiv u_{t-1} q_t,$$

so that the effective communication-induced reproduction factor becomes

$$r_{t,\mathrm{eff}}(d) = b\,(1 - m u_{t-1} q_t)^d\,.$$

The index $t-1$ reflects one-round delayed communication; an initially empty inbox contributes no communication pruning. The mean-field supply assumption corresponds to $u_{t-1} \simeq 1$. When $u_{t-1} < 1$, communication provides less pruning than predicted by mean field, shifting the critical degree toward

$$d_c^{\mathrm{empirical}} > d_c^{\mathrm{MF}}.$$

A.1.2 CORRELATED EVIDENCE

Even when experiment supply is sufficient, the assumption of independent neighbor evidence can fail at large communication degree. Different agents may receive strongly overlapping sets of evidence. Their exact hypothesis frontiers need not become similar, but their accumulated constraints can become correlated. Consequently, different neighbors may send duplicate or already implied evidence.

The possible feedback is summarized schematically as

$$\begin{aligned} d \uparrow \;&\Rightarrow\; \text{shared evidence} \uparrow \Rightarrow\; \text{constraint overlap} \uparrow \\ &\Rightarrow\; \text{redundant evidence} \uparrow \Rightarrow\; \text{communication efficiency} \downarrow . \end{aligned} \tag{10}$$

Duplicate observations alone do not establish this mechanism, since independent experiments can also coincide.

Let $s_{\mathrm{comm}}(t, d)$ be the survival probability of an incorrect candidate present immediately before communication. We define an effective communication degree through

$$s_{\mathrm{comm}}(t, d) = (1 - mq_t)^{d_{\mathrm{eff}}(t,d)} .$$

When the reference probability is accurate and neighbor pruning is approximately homogeneous and independent, $d_{\mathrm{eff}} \simeq d$. Redundancy can instead give $d_{\mathrm{eff}} < d$. This inequality is not universal: d_{eff} is a descriptive diagnostic and can also reflect inaccuracies in the reference probability or heterogeneity among candidates.

Holding the remaining reproduction dynamics fixed, weaker pruning can shift the candidate-growth crossing toward

$$d_c^{\mathrm{empirical}} > d_c^{\mathrm{MF}}.$$

A.2 PARENT-LOCKED REFERENCE

The parent-locked regime remains useful as an analytically tractable reference, but it should not be interpreted as the typical state of the synthetic system. In this limit all active agents share the same parent hypothesis at the beginning of a round,

$$H_t^{(1)} = H_t^{(2)} = \cdots = H_t^{(N)},$$

while independently sampling children from that common parent.

At depth t, the number of possible children is

$$S_t = K(M - t + 1).$$

If an active agent generates b children, the probability that a specific child h is generated is

$$P(h \in \mathcal{B}_t^{(j)}) = \frac{b}{K(M - t + 1)}.$$

Assume that the total experimental budget Nm can be fully used and redistributed among active agents. In the regime where its candidate capacity is sufficient to use this allocation,

$$P(h \text{ tested} \mid h \in \mathcal{B}_t^{(j)}, j \text{ active}) \simeq \frac{m}{a_t b}.$$

A randomly selected neighbor is active with probability a_t. Therefore, the pruning probability contributed by one nominal neighbor is

$$mq_t^{\mathrm{lock}} = a_t \frac{b}{K(M - t + 1)} \frac{m}{a_t b} = \frac{m}{K(M - t + 1)},$$

The active fraction and branching number cancel as long as the total experimental budget can indeed be used.

For $T \ll M$ and the standard full-depth approximation, this gives the scaling

$$d_c^{\mathrm{lock}} \simeq \frac{MK \ln b}{m} = \frac{|\mathcal{L}| \ln b}{m}.$$

The parent-locked limit should therefore be regarded as a reference regime in which agents share the same parent but retain independent child sampling.

In the concentrated regime, the critical communication degree is bounded by

$$\frac{2|\mathcal{L}|\ln b}{mT} \lesssim d_c^{\mathrm{conc}} \lesssim \frac{|\mathcal{L}|\ln b}{m}.$$

B Task Details

We consider three tasks: synthetic component search, TCAS configuration-fault localization, and physical mechanism discovery. All use the same branching-and-pruning protocol, but differ in the meaning of a hypothesis, the experimental interface, and the evidence available for pruning.

Throughout, $H^\star$ denotes the hidden mechanism, T its number of components, and h_t a hypothesis containing t components. Hypotheses are sets, so different refinement orders can reach the same node. The examples below are therefore partial views of a *confluent hypothesis graph*; blank nodes and ellipses indicate omitted hypotheses.

B.1 Synthetic Search Task

Hidden mechanism and hypotheses. The task has M candidate components, each with a value set $\mathcal{V}$ of size K. We write $\ell_{j,a}$ for the atomic assignment "component j takes value a," where $j \in \{1,\dots,M\}$ and $a \in \mathcal{V}$. The hidden mechanism is

$$H^\star = \{\ell_{j_1,a_1^\star},\dots,\ell_{j_T,a_T^\star}\}, \qquad T \le M,$$

with distinct component indices $j_1,\dots,j_T$. These indices are sampled uniformly from the size-T subsets of $\{1,\dots,M\}$, and their values are sampled uniformly from $\mathcal{V}$. A depth-t hypothesis contains t assignments to distinct components; there are $\binom{M}{t}K^t$ possible hypotheses at depth t. A partial hypothesis is truth-compatible exactly when $h_t \subseteq H^\star$.

Experiments and evidence. An experiment tests one proposed atomic assignment $\ell_{j,a}$. A refuted assignment prunes every local hypothesis containing it and is not regenerated; communicated experimental records allow neighbors to apply the same pruning rule. A run succeeds when at least one agent uniquely identifies $H^\star$ at depth T.

Success criterion. A run succeeds when at least one agent identifies the complete ground-truth mechanism $H^\star$ at depth T and its constituent assignments are supported by the accumulated experimental evidence.

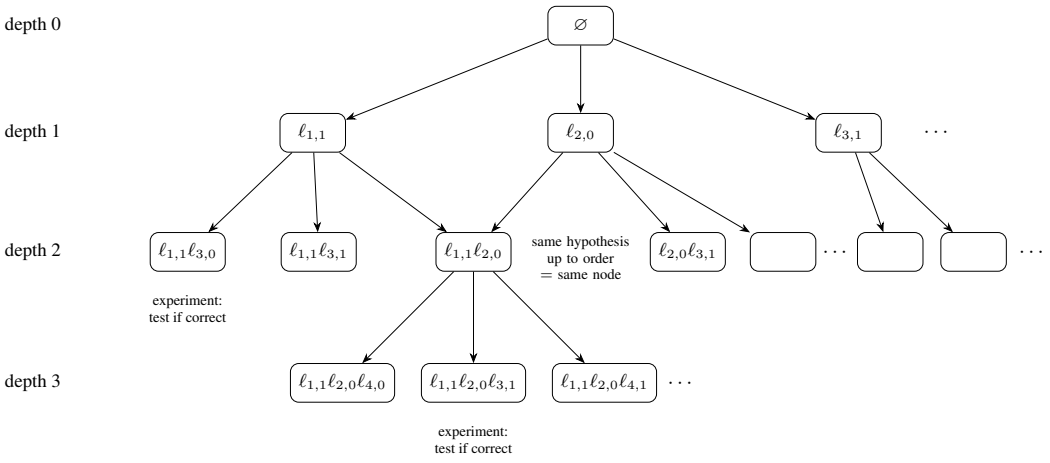


Figure 6: **Synthetic hypothesis search graph.** Each symbol $\ell_{j,a}$ denotes the atomic assignment that component j takes value a. Each refinement adds one previously unassigned component–value pair. Different assignment orders can reach the same hypothesis node, and experimentally refuted assignments prune all hypotheses that contain them.

B.2 TCAS CONFIGURATION-FAULT LOCALIZATION

Hidden mechanism and hypotheses. We use the configuration space of `tcas` from the Siemens test suite as a controlled fault-localization benchmark (Hutchins et al., 1994; Ghandehari et al., 2013). There are $M = 12$ categorical parameters. Parameter p_j has a legal value set $\mathcal{V}_j$ of size K_j. The parameter cardinalities are seven binary, two ternary, one four-valued, and two ten-valued, giving

$$|\mathcal{X}| = \prod_{j=1}^{12} K_j = 460{,}800, \qquad S = \sum_{j=1}^{12} K_j = 44.$$

Here $|\mathcal{X}|$ counts complete configurations, whereas S counts atomic parameter–value assignments. The hidden mechanism $H^\star$ is a failure-inducing combination of T assignments to distinct parameters. A depth-t hypothesis is

$$h_t = \{p_{j_1} = v_{a_1}, \ldots, p_{j_t} = v_{a_t}\},$$

and each refinement assigns one previously unassigned parameter. Different assignment orders can therefore reach the same partial configuration, as illustrated in Fig. 7.

Experiments and evidence. An experiment tests a complete configuration $x = (x_1, \ldots, x_M) \in \mathcal{X}$ compatible with a proposed refinement and returns $y(x) \in \{\text{pass}, \text{fail}\}$. We write $h \subseteq x$ when x satisfies every assignment in h. The configuration–outcome table can be precomputed, so an experimental query is implemented as a table lookup. Under the conjunctive fault model, a passing configuration rules out every candidate sufficient failure-inducing combination contained in it:

$$y(x) = \text{pass}, \quad h \subseteq x \quad \Longrightarrow \quad h \text{ is ruled out as a sufficient failure condition.}$$

A failing configuration does not identify which subset of assignments caused the failure and therefore does not, by itself, refute a particular atomic assignment. Agents communicate the records $(x, y(x))$, allowing neighbors to apply the same evidence to their own candidate hypotheses. Unlike the synthetic component test, this experiment therefore returns a verdict on a complete configuration rather than on one atomic assignment.

Success criterion. A complete candidate h predicts failure when $h \subseteq x$ and pass otherwise; denote this prediction by $\widehat{y}_h(x)$. It is evaluated on a separate held-out configuration set $\mathcal{E}_{\text{test}}$ using

$$D(h, H^\star) = \frac{1}{|\mathcal{E}_{\text{test}}|} \sum_{x \in \mathcal{E}_{\text{test}}} \mathbf{1}[\widehat{y}_h(x) \neq y^\star(x)],$$

where $y^\star(x)$ is the hidden system's outcome. A run succeeds when at least one agent reaches a depth-T hypothesis with $D(h, H^\star) < \epsilon$ that is also consistent with the configuration–outcome evidence accumulated during search.

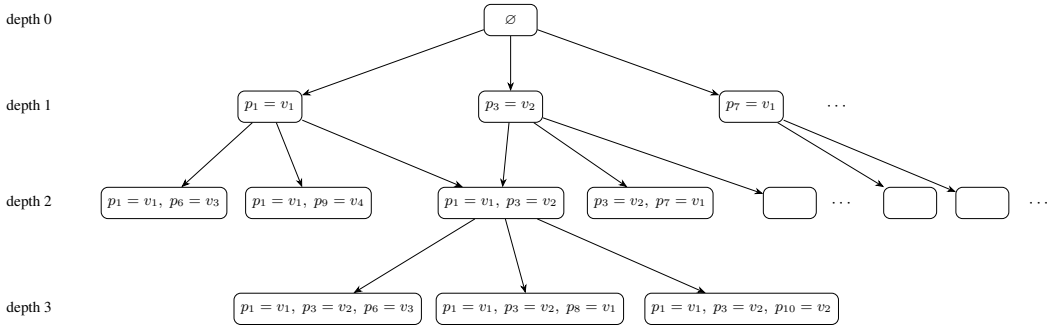


Figure 7: **TCAS configuration-fault hypothesis graph.** Each child adds one parameter–value assignment to its parent. Different assignment orders that produce the same partial configuration correspond to the same hypothesis node. Blank nodes and ellipses indicate additional hypotheses omitted for clarity.

B.3 PHYSICAL MECHANISM DISCOVERY TASK

Hidden mechanism and hypotheses. Following the dynamical-system discovery setting of ODEBench and the interactive probing setting of NewtonBench (d'Ascoli et al., 2024; Zheng et al., 2026), we consider a hidden equation of motion

$$\ddot{x} = f^\star(x, \dot{x}, \tau) = \sum_{c \in H^\star} c(x, \dot{x}, \tau), \qquad |H^\star| = T.$$

Here τ denotes physical time, to distinguish it from search depth t. Candidate terms come from M symbolic component families, each with K allowed parameter choices. An atomic component specifies both a symbolic family and its parameter choice. Examples include a linear restoring force, a cubic restoring force, linear or state-dependent damping, and periodic forcing. A depth-t hypothesis $h_t = \{c_1, \ldots, c_t\}$ contains t components, and a refinement adds one new symbolic component. Addition order does not affect the resulting mechanism, as illustrated in Fig. 8.

Experiments and evidence. Before the first communication round, each agent receives a different short trajectory from the same hidden system. These local observations probe different regions of state space and are held fixed across communication degrees. An experiment selects an initial state $z_0 = (x_0, \dot{x}_0)$ and starting time τ_0 from a fixed experiment bank and returns

$$\mathcal{E}(z_0, \tau_0) = \{(\tau_\ell, z(\tau_\ell))\}_{\ell=1}^{L}, \qquad \tau_\ell \in [\tau_0, \tau_0 + \Delta\tau],$$

where $z = (x, \dot{x})$. The observation duration $\Delta\tau$, number of measurements L, and measurement precision are fixed. Experiments are chosen to help distinguish a proposed component from the current hypothesis.

In the controlled rule-based version, experimental evidence is reduced to the idealized component-level outcome

$$y(c) = \mathbf{1}[c \in H^\star].$$

This is the stipulated evidence interface for the controlled experiment, not a claim that an arbitrary short trajectory uniquely identifies a term. Negative evidence removes every hypothesis containing c; positive evidence confirms c and excludes incompatible parameter choices in the same family. The testing agent keeps the physical trajectory private and communicates only the component-level record $(c, y(c))$.

Success criterion. A run succeeds when at least one agent discovers a depth-T mechanism that is consistent with the experimental evidence accumulated during search and predicts a separate set of held-out trajectories with error

$$D(h, H^\star) < \epsilon.$$

Here D denotes held-out trajectory prediction error, distinct from the configuration-classification error used for TCAS.

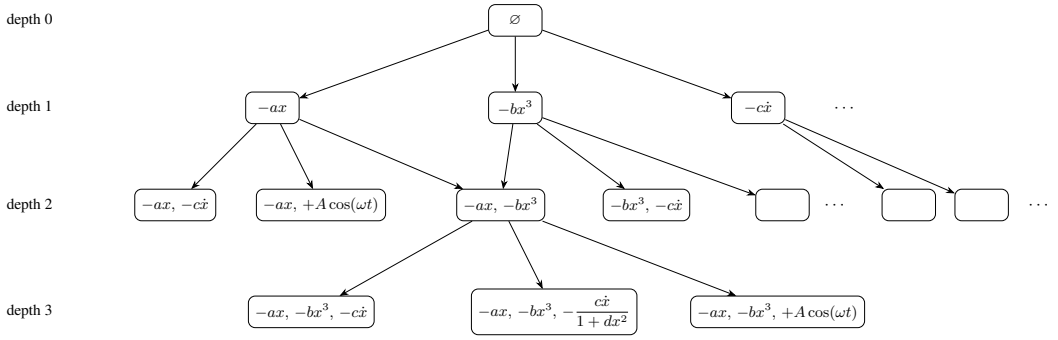


Figure 8: **Physical-mechanism hypothesis graph.** Each child adds one atomic symbolic component to its parent. Different addition orders can reach the same mechanism. Blank nodes and ellipses indicate additional hypotheses omitted for clarity.

C Additional Results

C.1 Sharpening of the Success Crossover

We test whether the success crossover becomes sharper as the search horizon T increases. We use the synthetic benchmark and vary only T, while fixing

$$N = 80, \qquad M = 8, \qquad K = 4, \qquad b = 3, \qquad m = 1.$$

For a given communication degree d, we define an episode as successful if *at least one agent recovers the true hypothesis with supporting evidence*. The empirical success probability is therefore

$$P_{\mathrm{succ}}(d) = \frac{1}{E} \sum_{e=1}^{E} \mathbf{1}[\text{at least one agent succeeds in episode } e], \tag{11}$$

where $E = 8$ episodes are evaluated at each degree. Consequently, the empirical probabilities take values in increments of $1/8$.

Figure 9 shows the raw empirical $P_{\mathrm{succ}}(d)$ for three representative horizons, $T = 3, 4, 5$. The crossover is broad at $T = 3$ and becomes progressively more concentrated in d as T increases.

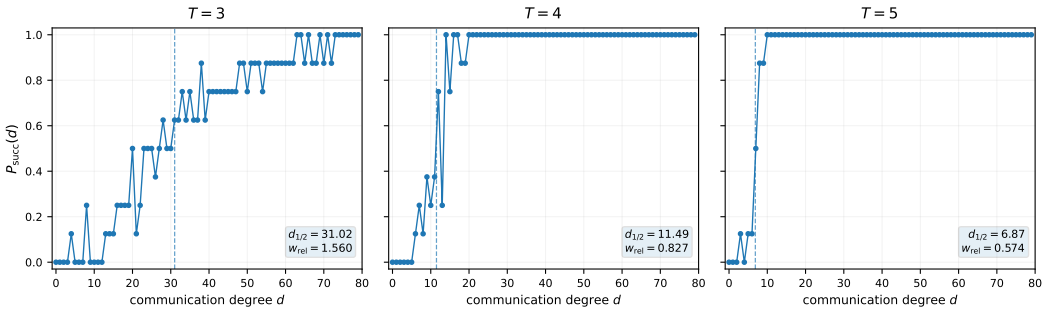


Figure 9: Empirical success probability $P_{\mathrm{succ}}(d)$ as a function of communication degree d for representative horizons $T = 3, 4, 5$ in the synthetic benchmark. Each point is the observed success fraction over eight independent task seeds, where an episode is successful if at least one agent recovers the ground-truth hypothesis with supporting evidence. Dashed vertical lines indicate the fitted crossover midpoint $d_{1/2}$. As T increases, the crossover shifts to lower communication degree and becomes visibly sharper.

To quantify this sharpening, we fit the episode-level binary success outcomes with a logistic crossover,

$$P_{\mathrm{fit}}(d) = \frac{1}{1 + \exp[-(d - d_{1/2})/w]}, \tag{12}$$

where $d_{1/2}$ is the degree at which the fitted success probability equals $1/2$, and w controls the crossover width. Let d_p denote the degree at which $P_{\mathrm{fit}}(d_p) = p$. We define the 10%–90% width as

$$\Delta d_{10\text{–}90} = d_{0.9} - d_{0.1} = w \ln 81.$$

Because the location of the crossover shifts substantially with T, we use the dimensionless relative width

$$w_{\mathrm{rel}} \equiv \frac{\Delta d_{10\text{–}90}}{d_{1/2}} = \frac{w \ln 81}{d_{1/2}}. \tag{13}$$

Smaller w_{rel} corresponds to a sharper crossover.

Across the full scan $T = 3, \ldots, 8$, we obtain

T	3	4	5	6	7	8
w_{rel}	1.560	0.827	0.574	0.443	0.436	0.069.

Table 3: Communication behavior under Team-Free and Indiv-Free policies, pooled over six task seeds and three environments.

Metric	Qwen3.5-4B		Qwen3.5-9B		GPT-5.4 Nano		GPT-6 Luna	
	Team-Free	Indiv-Free	Team-Free	Indiv-Free	Team-Free	Indiv-Free	Team-Free	Indiv-Free
Structured reports sent	6,702	6,921	5,081	4,975	8,965	8,556	9,103	9,004
Misreported outcomes	0.81%	0.87%	0.98%	1.19%	0.09%	0.09%	0.00%	0.00%
Sent nothing	13.8%	15.2%	27.4%	30.5%	13.3%	17.7%	5.7%	16.3%
Used free text	68.8%	63.2%	68.7%	64.8%	75.5%	64.9%	89.9%	70.1%
Positive findings disclosed	66.1%	65.6%	49.8%	49.0%	80.3%	75.5%	84.0%	82.5%
Negative findings disclosed	51.5%	53.2%	36.1%	35.4%	72.5%	69.1%	84.3%	82.4%

The relative width decreases strongly with T, with only a small finite-sample variation between $T = 6$ and $T = 7$.

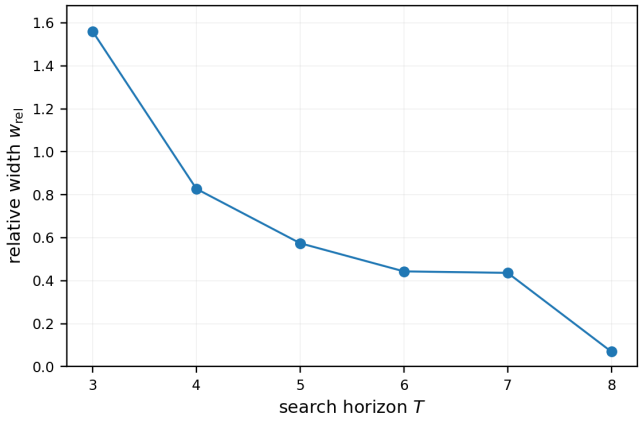


Figure 10: **Relative crossover width versus search horizon.** The relative 10%–90% width $w_{\mathrm{rel}} = (d_{0.9} - d_{0.1})/d_{1/2}$ obtained from logistic fits to the episode-level success outcomes. The width decreases strongly as T increases, indicating progressive sharpening of the finite-size success crossover. The line connecting the points is included only as a guide to the eye.

Together, Figs. 9 and 10 show that increasing the search horizon systematically sharpens the finite-horizon success crossover in communication degree. This progressive narrowing is consistent with increasingly phase-transition-like behavior as T grows, while the transition remains smooth at the finite horizons considered here.

C.2 LLM Communication Behavior under Free Communication

Table 3 shows a detailed breakdown of how agents chose to communicate under free communication policies Team-Free and Indiv-Free. Both conditions allow free-form messages, and the prompt explicitly permits agents to withhold a result or to report an outcome different from the one they observed. A very small percentage of Qwen agents misreported, but the most noticeable result is that Qwen models are generally more reticent than GPT models. Qwen models communicate positive verification results 49-66% of the time, while GPT models do 75-84% of the time. Qwen models are also more unlikely to disclose negative verification results (35-53%), compared to 69-84% by GPT models.

Table 4 shows d_c(ColumnName) $-$ d_c(Team-Auto): the difference between d_c for each free-commmunication policy and the baseline, automatic communication.

Table 4: Each entry is Δd_c, the paired difference in critical degree between a free-communication policy (Team-Free or Indiv-Free) and the fixed-communication baseline Team-Auto: positive means free communication *raises* d_c, so more communication is needed to suppress incorrect lineages. Intervals are exact 95% cluster bootstraps over those six seeds. Entries are $\Delta^{+\text{upper}}_{-\text{lower}}$.

Model	Synthetic		TCAS		Physics	
	Team-Free	Indiv-Free	Team-Free	Indiv-Free	Team-Free	Indiv-Free
Qwen3.5-4B	$6.27^{+0.48}_{-0.56}$	$6.92^{+2.55}_{-1.90}$	$4.90^{+6.77}_{-0.58}$	$7.33^{+4.16}_{-2.66}$	$5.17^{+1.09}_{-1.48}$	$5.54^{+2.88}_{-3.10}$
Qwen3.5-9B	$9.37^{+1.05}_{-2.03}$	$9.42^{+2.00}_{-2.08}$	$16.24^{+5.02}_{-5.47}$	$14.40^{+1.20}_{-1.99}$	$8.73^{+1.69}_{-2.18}$	$8.43^{+1.72}_{-2.13}$
GPT-5.4 Nano	$1.48^{+1.38}_{-0.84}$	$0.90^{+0.66}_{-0.63}$	$2.35^{+0.87}_{-1.62}$	$3.88^{+0.82}_{-3.16}$	$2.09^{+3.23}_{-1.97}$	$1.84^{+0.86}_{-1.19}$
GPT-6 Luna	$0.18^{+0.60}_{-0.57}$	$-0.42^{+0.67}_{-0.64}$	$3.66^{+3.51}_{-3.25}$	$3.08^{+0.96}_{-3.46}$	$0.32^{+0.84}_{-0.97}$	$-0.28^{+0.67}_{-0.54}$

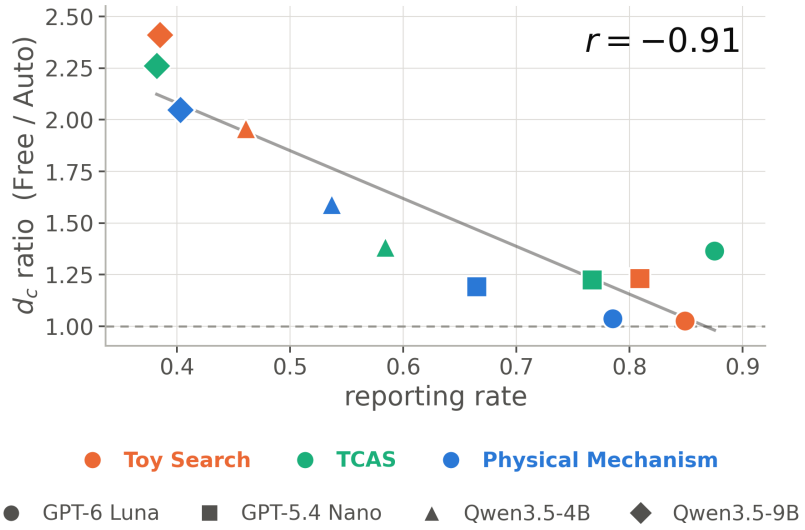


Figure 11: Correlation between ratio of Team-Free and Team-Auto d_c values and the rate at which models report their findings to their neighbors.

Figure 11 shows the correlation between the ratio of d_c measured from Team-Free and Team-Auto policies (representing the difference in d_c that perfect communication makes) and the rate at which models report their findings to their neighbors.

C.3 Effect of Increasing Reasoning Effort

Figure 12 shows a nuance to the story that increased reasoning effort leads to worse results (higher d_c). While d_c increases and deviates further from theory with more reasoning, reasoning often improves agents' ability to solve the task (especially on physical mechanism discovery task). This effect is true at higher communication degrees as well, suggesting that communication still helps.

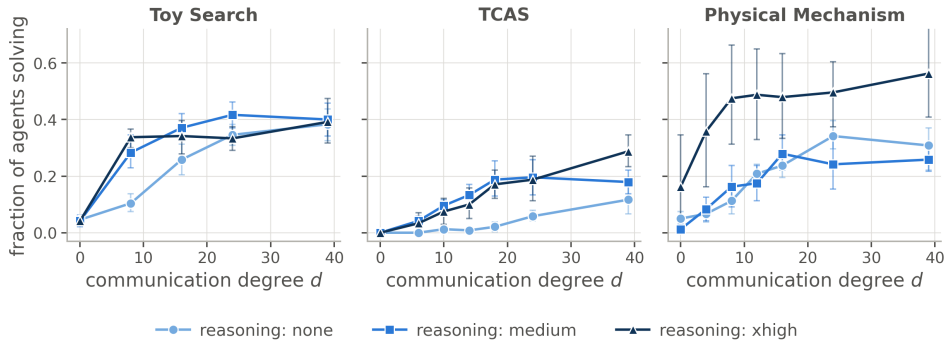


Figure 12: Fraction of agents completing the search increases with reasoning in GPT-6-luna.