The Cost of Coordination: Why Multi-Agent LLMs Explore More But Discover Less

Shirley S. Qiu shirleyqiu@ucsb.edu
Mable M. Zhou mzhou@ucsb.edu
Jonathan W. Schooler jschooler@ucsb.edu

Department of Psychological and Brain Sciences,
University of California, Santa Barbara

Abstract

Large language models (LLMs) are increasingly deployed in multi-agent teams (MAS), yet recent works have shown that unstructured multi-agent collaboration often degrades performance in noninsight problems. We examined whether previous findings can be generalized to an insight problem solving paradigm and identified process-level signatures that may explain the performance gap. We compared performance between solo versus two-agent ChatGPT-4o teams to solve situation puzzles in a 2 (Agent Configuration) × 2 (Solving time) experiment (N = 82). Teams were significantly less accurate than solo agents despite asking questions across broad categories. However, question quality was equivalent between conditions. A multilevel mediation analysis controlling for time decomposed the coordination cost into two components: a protocol failure component (47%), in which inter-agent discussion reduced host-directed questioning, and a residual coordination cost (53%) consistent with impaired convergent integration that persisted after controlling for question volume and quality. Doubling the time budget proportionally increased question volume in both conditions but preserved the team to solo ratio, indicating that the cost is structural rather than driven by limited time. These findings extend multi-agent coordination costs to insight problem solving and suggest that unstructured multi-agent interaction supports the divergent phase but disrupts the convergent phase.

Keywords: Multi-Agent System (MAS); Large Language Models (LLMs); Process Losses; Problem Solving; Group Cognition

Introduction

With the increasing use of large language models (LLMs) across diverse settings, and with the long-term goal of reaching artificial general intelligence, LLMs agents will inevitably need to work in teams and coordinate with one another to solve complex problems (Gottweis et al., 2025; Fourney et al., 2024). It is interesting and important to understand how LLMs teams solve problems, and whether team-based interaction facilitates or inhibits performance, particularly when solving novel complex problems through conversation compared to single agent working independently.

Multi-agent systems (MAS) have been proposed as a way to amplify LLM capacities by exploiting exploration and interaction among agents (Guo et al., 2024; Qian et al., 2024; Liang et al., 2024). However, a growing body of evidence has challenged this assumption. Previous works showed that increase in agents' network size reduces performance, with unstructured collaboration producing the worst outcomes compared to other multi-agent configurations (Grötschla et al., 2025; Radeva et al., 2025). These findings establish multi-agent degradation as a robust empirical phenomenon across tasks and architectures. However, the existing literature is limited in two aspects. First, prior works have focused primarily on logical problem solving tasks, where solutions can be reached through incremental deduction. It remains an open question whether the coordination cost generalizes to insight problem solving, where the task structure may differentially reward broad exploration that multi-agent interaction tends to produce. Second, previous works have documented the team underperformance comparing to solo agents without investigating how the multi-agent interaction shapes the problem solving process. Understanding these process-level dynamics is essential for diagnosing why multi-agent teams underperform, and whether such cost is inherent to coordination or specific to particular phases of problem solving.

While solving the logical problems often depend on the accumulation of information, solving insight problems often require a shift in representing the problem framing and its constraints. Ohlsson's Representational Change Theory formalizes this process: solvers reach an impasse when they hold a dominant but inappropriate encoding of a problem, and they must relax problem constraints and re-interpret the problem framing in order to solve the problem, which gives rise to insight (Ohlsson, 1992; Knoblich et al., 1999). It is worth noting that solving an insight problem does not solely rely on divergent exploration of potential solutions – the solver must also iteratively test hypotheses under the new framing to converge on the ground truth (Cropley, 2006). This establishes the dual nature of insight problems, which is underpinned by both divergent representation-restructuring and convergent hypothesis-testing. Existing benchmarks tested on whether LLMs can solve insight problems, including the *BRAINTEASER* benchmark, which challenges models to defy common sense expectations in word and sentence puzzles (Jiang et al., 2023), and *LatEval*, which evaluates solo LLMs on interactive lateral thinking puzzles where they need to find an unexpected solution to a seemingly straightforward puzzle (Huang et al., 2024). Both benchmarks reveal that single LLM agents are struggling with insight problems, and thus achieve a performance worse than human baseline.

To study how coordination reshapes the process and outcome of insight problem solving, we employed the situation puzzles. Situation puzzles provide the solvers with a short and ambiguous scenario, requiring the solver to reconstruct

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

the unique hidden story that explains the scenario by asking the host yes/no questions. Solving a situation puzzle requires both divergent and convergent thinking: the solver must restructure their mental representation by reframing the problem and relaxing initial constraints, then systematically test new hypotheses to converge on the ground truth. For example, a puzzle may describe "*a man who walks into a bar, is greeted by a bartender pointing a gun, says 'thank you', and walks out*". The surface description activates an intuitive but incorrect framing in which the gun is a threat. Solving the puzzle requires relaxing this constraint and discovering that the man entered to cure his hiccups and that the gun served to startle him out of them. The puzzle cannot be solved through divergent leaps alone, nor through logical deduction within the wrong problem space; neither process is sufficient on its own to solve the puzzle.

Critically, situation puzzles externalize the sequential reasoning trajectory in a way that classical insight paradigms do not. In classical insight tasks, such as the matchstick arithmetic task, where solvers must transform a false Roman-numeral equation (e.g., IV = III + III) into a true one by moving only one matchstick (Knoblich et al., 1999), or the Remote Associates Task, where solvers must identify a single word that forms phrases with three seemingly unrelated cue words (e.g., cream skate cube $\rightarrow$ ice) (Mednick, 1962), the process of reconstructing the problem and its constraints is largely internal, with the outcome evaluated on a binary scale as either solved or unsolved. Concurrent verbal protocols can partially externalize this process, but their quality depends on solvers' ability to articulate their reasoning, and verbalization itself can disrupt insight (Schooler et al., 1993). By contrast, rather than relying on self report, the situation puzzle task can externalize the reasoning chain with its task structure. Each yes/no question serves as an observable search in the problem space, with immediate binary feedback enabling the solver to update their hypotheses. Because each question depends on the accumulated conversation history, the task approximates a sequential hypothesis-testing process in which solvers re-encode problem elements, relax constraints, and narrow the hypothesis space across turns. Solving the puzzle therefore depends on chaining information into a coherent story rather than generating a single clever question, and the emergence of insight can be externally traced through the question trajectory rather than inferred from a single outcome measure.

The present study compares the performance of solo LLM agents against unstructured two-agent LLMs teams on Situation Puzzles Task, a paradigm that involves divergent-convergent thinking process and externalizes the reasoning trajectory through asking yes/no questions sequentially. We tested whether the multi-agent coordination cost generalizes to insight problem solving. By leveraging the process-tracing structure of the Situation Puzzle Task, we explored the mechanisms underlying the performance gap by tracking question quality and exploration diversity. We also examined whether time constraints modulated the coordination cost, since inter-agent interaction itself consumes time that could otherwise be directed toward hypothesis-testing. Because situation puzzles reward the broad exploration that team interaction tends to generate, any team disadvantage on this task would indicate a coordination cost that persists even under a favorable task structure. We predicted that unstructured two-agent LLM teams would underperform solo LLM agents while exhibiting greater exploratory diversity in their question trajectories. We further hypothesized an interaction effect between agent configuration and time constraints: allowing longer time to solve the puzzle will compensate for the coordination cost of agent teams and narrow their performance gap with the solo agent.

Method

We employed a 2 (Agent configuration: solo vs. paired team) × 2 (Time constraint: 30-minute vs. 1-hour) between-subjects factorial design with N = 82 independent trials. The 30-minute condition included 50 trials (25 solo, 25 team), and the 1-hour condition included 32 trials (16 solo, 16 team). Each trial involved either a single GPT-4o agent or an unstructured paired GPT-4o team attempting to solve 3 situation puzzles under their given time constraint.

Situation Puzzle Task

Situation puzzles, also known as lateral thinking puzzles, present an apparently paradoxical scenario requiring solvers to uncover a hidden narrative through yes/no questions. The host will give the ending of the scenario to the solver, and the solver's task is to recover the specific plot that led to the ending by asking the host Yes-or-no questions. The host will only respond with yes or no to the question until the plot is fully resolved. Each puzzle contains "key elements" that (1) require insight to reveal hidden plot elements and (2) elicit an "aha" moment, which cannot be reached through deduction alone. Such elements may be protagonist identity, location, time, causality, or supernatural elements. The task structure creates diagnostic sensitivity to coordination costs: solvers must first generate a divergent insight to identify the problem space, then conduct convergent hypothesis testing to systematically narrow possibilities and reconstruct plot elements.

The Situation Puzzle Task is a 30-minute or 1-hour web-based task comprising 3 distinct puzzles. Each puzzle features 3 different key elements to minimize practice effects. All the puzzles are designed by the experimenters and none of the puzzles are drawn from publicly available datasets accessible to LLMs, reducing the risk of data contamination.

The solver, played by ChatGPT-4o model, may verify their solution by asking the host, played by Gemini-2.0-flash model, whether their current plot is correct. Once confirmed, the puzzle ends and the team proceeds to the next puzzle. The solver may choose to switch to a new puzzle after 10 questions. If the solvers are paired, they need to reach an agreement before choosing to switch puzzles. Switched puzzles cannot be revisited.

Instruction to LLM agents

ChatGPT-4o as Problem Solver. All solver agents used GPT-4o configured with temperature = 0.7, top-p = 1.0. No seed parameter was specified, allowing for natural response variability across trials. The context window size was 128,000 tokens. The agents have access to full conversation history. The system prompt for agents in both solo and team conditions included explicit decision logic for: (1) internal discussion between agents, (2) asking the Host questions, (3) proposing solutions, and (4) giving up. Agents in team condition received no role differentiation, task decomposition guidance, or consensus mechanisms beyond the shared conversation context. This unstructured protocol was designed to maximize coordination overhead by requiring agents to self-organize without explicit coordination scaffolding.

Gemini-2.0-flash as the Host. To maintain experimental consistency, all yes/no question answering and solution evaluation was conducted by a single automated host using Google gemini-2.0-flash model. No explicit temperature, top-p, or seed parameters were configured, relying on model defaults (temperature = 1.0, top-p = 0.95).

The host remained constant across all experimental conditions with the same prompt given. The structured prompt contains: (1) the full puzzle solution that is unknown to the solvers, (2) the puzzle clue that is known to the solvers, (3) the list of key plot elements to track, and (4) the agent's question. The host was instructed to (1) determine whether the question warrants "Yes," "No," "Irrelevant", (2) estimate a warmness score (0-100) representing how close the question or reasoning was to the core solution, with higher scores indicating greater proximity to key plot elements, (3) identify whether the question or answer revealed any of the predefined key plot elements, returning the amount of key element founded, and (4) evaluate whether the proposed solution accurately described key events and causality.

Dependent Measures

Performance score evaluation. The performance score is calculated by the cumulative number of plot elements found across all 3 puzzles from the task. The plot elements are pre-defined components that represents essential hidden information about the plot (e.g. "The man is playing Monopoly instead of being bankrupted in reality.") . There are 3 plot elements per puzzle. If the problem solver identified all 3 elements, they are treated as successfully solve the entire puzzle. A plot element discovered is counted whenever the LLM host flags that the solver's question or proposed answer has successfully revealed one or more of these predefined key elements.

Question Category Classification. To quantify exploratory diversity, each question was classified into one of six categories based on the primary information type sought (Table 1). Classification was performed via keyword matching using a deterministic algorithm. This approach ensures perfect reproducibility and eliminates subjective coding bias.

Table 1: Question Category Classification Scheme

Category	Description
AGENT	People, identity (e.g., who, alive)
CONTEXT	Location, setting (e.g., where, place)
ACTION	Events, behaviors (e.g., did, happen)
OBJECT	Physical items (e.g., weapon, car)
META	Puzzle structure (e.g., story, riddle)
UNKNOWN	Unmatched patterns

Entropy Calculation. The entropy is a quantitative measure towards exploration diversity. For each trial, we computed Shannon entropy of the question category distribution:

$$H = -\sum_{i=1}^{6} p_i \log_2(p_i)$$

where p_i represents the proportion of questions in category i. Entropy ranges from 0, representing all questions in one category, to $\log_2(6) \approx 2.58$, representing questions evenly distributed across all categories. Higher entropy indicates greater distributional diversity across question types, reflecting less focused search patterns. If a trial contained no valid questions or all categories were undefined, entropy was set to 0.

Warmness Calculation. Warmness is estimated by the LLM host for each question asked. The host, which has access to the full puzzle solution, evaluates the semantic proximity of each question to the core plot and key elements, assigning a score from 0 to 100. This scoring is based on the host model's in-context judgment of how close the question's content is to revealing the hidden narrative.

Results

We examined the effects of agent configuration (solo vs. team) and time constraint (30 minutes vs. 1 hour) on three dependent measures: (1) plot elements discovered (performance), (2) entropy of question trajectories (exploration diversity), and (3) warmness scores (semantic proximity of questions to the solution). All analyses used linear mixed-effects models with puzzle as a random intercept to account for variability in puzzle difficulty, fitted using lme4 in R. All analyses used $\alpha = .05$. Descriptive statistics are reported as session-level summaries across three puzzles.

Performance, Exploration Diversity, and Question Quality

Performance. Solo agents discovered significantly more plot elements ($\bar{x} = 3.59$, $SD = 2.02$) than team agents ($\bar{x} = 1.48$, $SD = 1.42$), with the mixed-effects model confirmed this effect ($\beta = -0.985$, $SE = 0.196$, $t = -5.03$, $p < .001$). Neither the main effect of time constraint ($\beta = -0.202$, $SE =$

0.190, $t = -1.06$, $p = .290$) nor the interaction between agent configuration and time constraint ($\beta = 0.247$, $SE = 0.284$, $t = 0.87$, $p = .386$) was significant, indicating that the solo agent performed better than the team regardless of time limits (30-minute: $\bar{x}_{\text{solo}} = 2.96$, $\bar{x}_{\text{team}} = 1.20$; 1-hour: $\bar{x}_{\text{solo}} = 4.50$, $\bar{x}_{\text{team}} = 1.81$) (Figure 1).

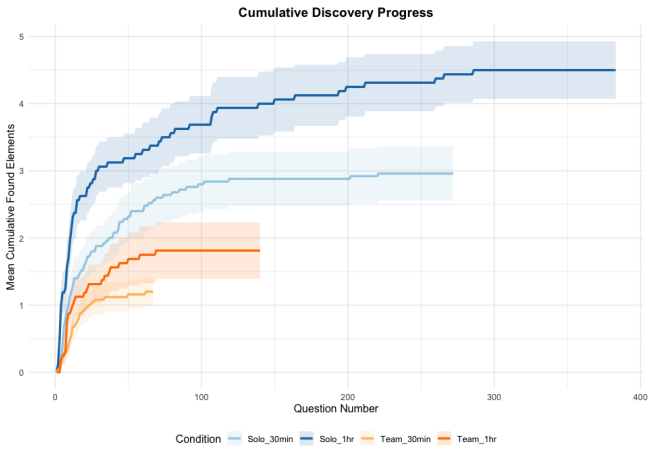

Figure 1: Cumulative discovery progress by condition. The lines represent the mean total elements found across all puzzles by question number. Solo agents (blue) exhibited steeper early discovery slopes and sustained information gain over time, whereas teams (orange) saturated earlier with fewer total questions.

Question Category Classification. Across all trials, agents used an average of 4.95 categories per session (Solo: $\bar{x} = 4.71$, $SD = 1.21$; Team: $\bar{x} = 5.20$, $SD = 1.03$) (Figure 2). Among the six categories, the identity of "character" (coded as "AGENT") were asked the most frequently (53.5% of all questions), followed by ACTION related questions (23.9%) and OBJECT (9.4%). CONTEXT questions were rare (2.9%), while META (3.5%) and UNKNOWN (6.7%) questions comprised a small proportion of the total. Solo agents predominantly relied on AGENT questions, which appeared as the most common category in 30 of 41 sessions. By contrast, teams distributed their focus more broadly, with AGENT dominating in only 19 of 41 sessions and META questions appearing as the primary category in 11 team sessions.

Exploration Diversity. Teams exhibited significantly higher entropy ($\bar{x} = 1.92$, $SD = 0.40$) than solo agents ($\bar{x} = 1.24$, $SD = 0.54$), confirmed by the mixed-effects model ($\beta = 0.556$, $SE = 0.110$, $t = 5.04$, $p < .001$) and indicated that teams explored more diverse question categories. The main effect of time constraint was not significant ($\beta = -0.063$, $SE = 0.107$, $t = -0.58$, $p = .561$), nor was the interaction ($\beta = 0.276$, $SE = 0.161$, $t = 1.72$, $p = .088$), indicating that the greater exploratory diversity in teams was consistent across time conditions (Figure 3).

Question Quality. A mixed-effects model on warmness scores revealed no significant main effect of agent configuration ($\beta = -0.878$, $SE = 0.964$, $t = -0.91$, $p = .364$), time constraint ($\beta = -1.260$, $SE = 0.936$, $t = -1.35$, $p = .180$), or their interaction ($\beta = 2.142$, $SE = 1.401$, $t = 1.53$, $p = .128$). Across conditions, mean warmness remained statistically equivalent between solo agents and teams, indicating comparable semantic quality of questions regardless of agent configuration.

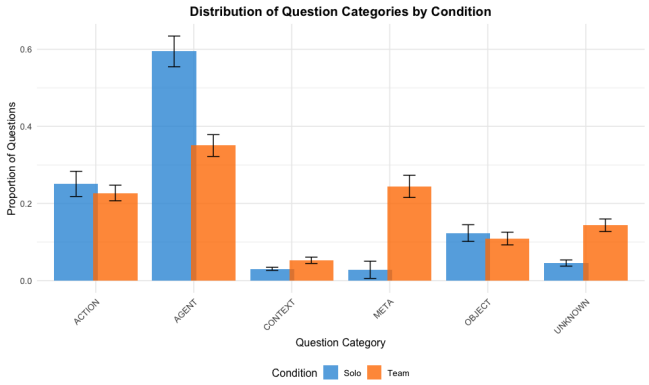

Figure 2: Distribution of question categories by condition. Solo agents (blue) concentrated questions in ACTION and AGENT categories, whereas teams (orange) distributed questions more evenly across all six categories with a concentration on AGENT and META categories. Error bars represent standard errors.

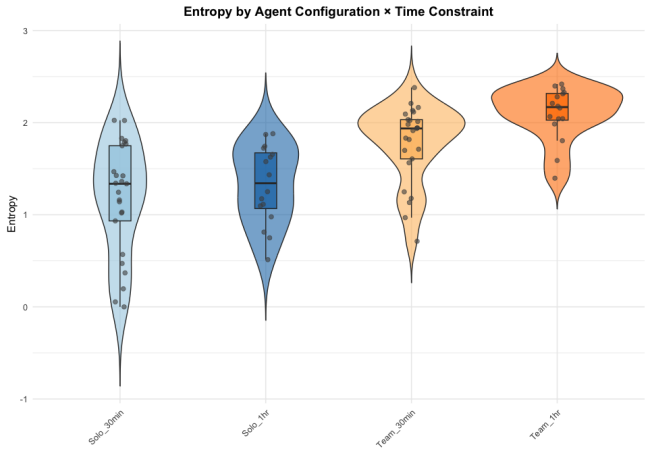

Figure 3: Entropy distributions by agent configuration and time constraint. Violin plots show entropy values for solo agents (blue) and team agents (orange) across different time conditions (30-minute, 1-hour). Box plots within violins display median, interquartile range, and individual data points.

Search Metrics and Performance

To assess whether exploration diversity and question quality independently predicted performance, we fitted a mixed-

Table 2: Total questions per session by agent configuration and time constraint.

Configuration	Time	*n*	*M*	*SD*
Solo	30 min	25	179.9	56.8
Solo	1 hour	16	374.8	91.7
Team	30 min	25	49.2	11.7
Team	1 hour	16	95.9	27.8

effects model with found elements as the outcome, mean warmness, entropy, agent configuration, and their two-way interactions (warmness × condition; entropy × condition) as fixed effects, with puzzle as a random intercept.

Warmness significantly predicted found elements (β = 0.060, SE = 0.017, t = 3.52, $p < .001$), indicating that questions with greater semantic proximity to the solution were associated with greater plot element discovery. By contrast, entropy did not significantly predict found elements (β = 0.077, SE = 0.164, t = 0.47, p = .639), suggesting that broader exploration across question categories did not independently contribute to discovery once warmness and condition were controlled. Neither the warmness × condition interaction (β = 0.052, SE = 0.032, t = 1.64, p = .102) nor the entropy × condition interaction (β = −0.092, SE = 0.227, t = −0.41, p = .684) was significant, indicating that the predictive relationships did not differ between solo and team agents. The agent configuration effect remained significant after controlling for both search metrics (β = −1.492, SE = 0.493, t = −3.02, p = .003).

These results suggest that despite teams' higher entropy indicating broader hypothesis exploration, such breadth does not translate into performance. Instead, performance is mainly driven by question quality of individual questions. The equivalence in question quality between solo and team agents, and the persistent differences in performance after controlling for both search metrics, together suggest the existence of a coordination cost that is not captured by either question quality or exploration diversity.

Mediation Analysis

Given that neither exploration diversity nor question quality explained the performance gap, we considered whether the total volume of host-directed questions could account for the deficit. Table 2 shows that teams directed substantially fewer turns toward the host across both time conditions (approximately 26-27% of the solo total), suggesting that inter-agent discussion displaced hypothesis-testing turns.

To formally test whether differences in total questions mediated the performance gap, with mean warmness retained as a parallel mediator, we estimated a parallel mediation model with total questions (M_1) and mean warmness (M_2) as candidate mediators, time condition as a covariate on all paths, and puzzle as a random intercept. All continuous variables were z-scored prior to analysis to place path coefficients on a standardized scale. Agent configuration remained binary (Solo = 0, Team = 1). Bootstrap confidence intervals for the indirect effects were computed using 5,000 BCa iterations.

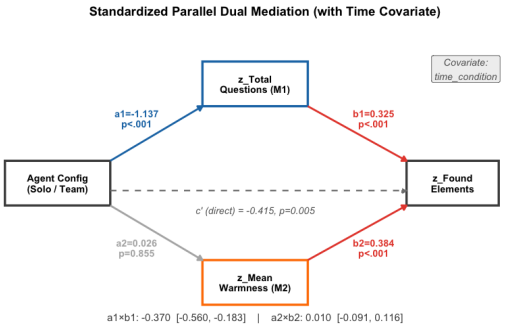

Figure 4: Standardized parallel mediation with time covariate: Condition → [Total Questions + Mean Warmness] → Found Elements. All continuous variables were z-scored; 5,000 BCa bootstrap iterations. The total-question pathway was significant (indirect = −0.370, 95% BCa CI [−0.560, −0.183]); the warmness pathway was not (indirect = 0.010, 95% BCa CI [−0.091, 0.116]).

The total-question pathway (M_1) was significant. Agent configuration strongly predicted total questions (a_1 = −1.137, SE = 0.118, t = −9.66, $p < .001$), and total questions significantly predicted found elements after controlling for condition, warmness, and time (b_1 = 0.325, SE = 0.080, t = 4.06, $p < .001$). The indirect effect through total questions was significant ($a_1 \times b_1$ = −0.370, 95% BCa CI [−0.560, −0.183]; Sobel z = −4.06, $p < .001$), accounting for approximately 47% of the total effect (c = −0.776). This indicates that reduced host-directed questioning in teams accounts for roughly half of the performance deficit, consistent with inter-agent discussion displacing hypothesis-testing turns.

The mean warmness pathway (M_2) was not significant. After controlling for time, agent configuration did not predict mean warmness (a_2 = 0.026, SE = 0.140, t = 0.18, p = .85), despite warmness significantly predicting found elements after controlling for condition and total questions (b_2 = 0.384, SE = 0.067, t = 5.73, $p < .001$). The null a-path indicated that agent configuration did not differentially affect question quality. Therefore, warmness did not mediate the group difference in found elements, despite being a significant predictor of performance.

The direct effect of agent configuration remained significant after controlling for both mediators (c' = −0.415, SE = 0.146, t = −2.84, p = .005), representing approximately 53% of the total effect. This pattern of partial mediation indicates that question displacement is not the sole mechanism through which coordination disrupts performance, as there is a substantial residual coordination cost persisting after equating conditions on question volume, question quality, and time (Figure 4).

Discussion

The present study extended the coordination cost previously revealed in logical problem solving tasks (Grötschla et al., 2025; Radeva et al., 2025) to insight problems. Across time constraints, solo LLM agents consistently outperformed paired LLM teams in plot element discovery, with the mixed effects model confirming a significant performance gap that was independent of time. The present results further reveal a dissociation: teams explored more broadly than solo agents yet discovered fewer plot elements while their question quality remained equivalent. Such pattern indicates that unstructured multi-agent interaction differentially affects the divergent and convergent processes of insight problem solving.

To localize the source of the coordination cost, we used a parallel mediation analysis controlling for time to test whether reduced host-directed questioning could fully account for the team performance deficit. The mediated component, transmitted through reduced host-directed questioning, accounted for approximately 47% of the team performance deficit. This aligns with previous work attributing the performance gap to protocol failure, in which inter-agent coordination displaces task directed effort (Radeva et al., 2025). However, there is still a residual coordination cost (53%) that persisted after equating conditions on question volume, question quality, and time, indicating that protocol failure is not the sole mechanism through which unstructured coordination disrupts performance. We thus infer this residual may arise from an impairment in convergent integration.

Two converging observations support this interpretation. First, teams exhibited significantly broader exploration in question categories than solo agents, but entropy did not independently predict the elements found after controlling for the volume and quality of the questions. Broader exploration therefore did not translate into more discoveries: teams searched diversely but found less. Second, cumulative discovery curves (Figure 1) show teams plateauing earlier than solo agents despite comparable question quality, indicating reduced information gain per turn. Together, these observations situate the residual coordination cost at the convergent stage of insight problem solving, where solvers must chain accumulated feedback into a narrowing reconstruction of the plot (Cropley, 2006). Multi-agent interaction thus appears to support divergent expansion of the hypothesis space while disrupting the convergent process that is crucial for translating accumulated information gain into solution.

The time manipulation provides further evidence that this coordination cost is structural rather than a transient consequence of time scarcity. Doubling the time budget scaled question volume proportionally in both conditions (Solo: 179.9 to 374.8; Team: 49.2 to 95.9), but the team-to-solo ratio remained relatively stable (~26-27%). The displacement of host directed questions is therefore not a time budget problem that additional turns can fix, but a stable property of how unstructured multi agent interaction allocates turns between coordination and question-asking.

The present findings have implications for how coordination costs are theorized across agent types. Prior research on human groups has documented reliable reductions in collective output under unstructured collaboration, typically attributed to production blocking, in which capacity constraints such as working memory limitations during turn taking force agents to abandon ideas while waiting for the floor (Diehl & Stroebe, 1987). The present findings show that structurally similar costs emerge in LLM teams despite the absence of analogous capacity constraints, as GPT-4o agents retain full conversation history and face no memory decay. This dissociation between capacity constraints and coordination cost suggests that the cost of unstructured coordination is not solely a product of cognitive capacity limits but may reflect a more general property of how interaction structure allocates effort between coordination and task directed action. Resolving this cost will likely require structural interventions, such as explicit role differentiation, structured turn taking, or scaffolds that preserve convergent integration.

Limitations

Two limitations qualify the present interpretation. First, although the residual coordination cost (53%) is consistent with an impairment in convergent integration, this interpretation rests on residual variance after controlling for question volume, quality, and time, rather than on direct measurement of integration efficiency. Second, the cross agent parallel with human production blocking remains an analogy in the absence of empirical comparison.

Future Directions

Two directions for future work follow directly. First, process-level analyses of agent interactions, including turn by turn dialogue patterns, hypothesis transitions, and cross turn information uptake, would test the integration deficit interpretation by converting residual based inference into direct measurement. Second, comparing matched human and LLM teams on the same paradigm would clarify whether the present coordination costs reflect analogous mechanisms across agents.

Conclusion

The present findings establish that multi-agent coordination costs extend to insight problem solving and decompose into two components: a protocol failure component (47%), in which inter-agent discussion reduces host-directed questioning, and a residual cost (53%) consistent with impaired convergent integration after controlling for question volume, quality, and time. The team disadvantage cannot be attributed to narrower exploration or worse individual questions: teams explored more broadly than solo agents, and question quality was equivalent across conditions. Achieving collective intelligence in multi-agent systems will require coordination structures that preserve convergent integration.

Acknowledgments

Authors used AI tools in preparing experiment materials and preparing their manuscripts. This research was supported by the UCSB Undergraduate Research and Creative Activities grant awarded to Shirley S. Qiu. We thank the two anonymous reviewers for their feedback.

References

Cropley, A. (2006). In praise of convergent thinking. *Creativity Research Journal, 18*(3), 391–404. https://doi.org/10.1207/s15326934crj1803_3

Diehl, M., & Stroebe, W. (1987). Productivity loss in brainstorming groups: Toward the solution of a riddle. *Journal of Personality and Social Psychology, 53*(3), 497–509. https://doi.org/10.1037/0022-3514.53.3.497

Fourney, A. C., Bansal, G., Mozannar, H., Tan, C., Salinas, E., Zhu, E., Niedtner, F., Proebsting, G., Bassman, G., Gerrits, J., Alber, J., Chang, P., Loynd, R., West, R., Dibia, V., Awadallah, A., Kamar, E., Hosn, R., & Amershi, S. (2024). Magentic-One: A generalist multi-agent system for solving complex tasks. *arXiv*. https://arxiv.org/abs/2411.04468

Gottweis, J., Weng, W.-H., Daryin, A., Tu, T., Palepu, A., Sirkovic, P., Myaskovsky, A., Weissenberger, F., Rong, K., Tanno, R., Saab, K., Popovici, D., Blum, J., Zhang, F., Chou, K., Hassidim, A., Gokturk, B., Vahdat, A., Kohli, P., . . . Natarajan, V. (2025). Towards an AI co-scientist. *arXiv*. https://arxiv.org/abs/2502.18864

Grötschla, F., Müller, L., Tönshoff, J., Galkin, M., & Perozzi, B. (2025). AgentsNet: Coordination and collaborative reasoning in multi-agent LLMs. *arXiv*. https://arxiv.org/abs/2507.08616

Guo, T., Chen, X., Wang, Y., Chang, R., Pei, S., Chawla, N. V., Wiest, O., & Zhang, X. (2024). Large language model based multi-agents: A survey of progress and challenges. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence (IJCAI 2024)* (pp. 8048–8057). International Joint Conferences on Artificial Intelligence Organization. https://doi.org/10.24963/ijcai.2024/890

Huang, S., Ma, S., Li, Y., Huang, M., Zou, W., Zhang, W., & Zheng, H. (2024). LatEval: An interactive LLMs evaluation benchmark with incomplete information from lateral thinking puzzles. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 10186–10197). ELRA and ICCL. https://aclanthology.org/2024.lrec-main.889/

Jiang, Y., Ilievski, F., Ma, K., & Sourati, Z. (2023). BRAINTEASER: Lateral thinking puzzles for large language models. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 14317–14332). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.885

Knoblich, G., Ohlsson, S., Haider, H., & Rhenius, D. (1999). Constraint relaxation and chunk decomposition in insight problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*(6), 1534–1555. https://doi.org/10.1037/0278-7393.25.6.1534

Liang, T., He, Z., Jiao, W., Wang, X., Wang, Y., Wang, R., Yang, Y., Shi, S., & Tu, Z. (2024). Encouraging divergent thinking in large language models through multi-agent debate. *arXiv*. https://arxiv.org/abs/2305.19118

Mednick, S. A. (1962). The associative basis of the creative process. *Psychological Review, 69*(3), 220–232. https://doi.org/10.1037/h0048850

Ohlsson, S. (1992). Information-processing explanations of insight and related phenomena. In M. T. Keane & K. J. Gilhooly (Eds.), *Advances in the psychology of thinking* (Vol. 1, pp. 1–44). Harvester-Wheatsheaf.

Qian, C., Xie, Z., Wang, Y., Liu, W., Zhu, K., Xia, H., Dang, Y., Du, Z., Chen, W., Yang, C., Liu, Z., & Sun, M. (2025). Scaling large language model-based multi-agent collaboration. *arXiv*. https://doi.org/10.48550/arXiv.2406.07155

Radeva, I., Popchev, I., Doukovska, L., & Dimitrova, M. (2025). Multi-agent coordination strategies vs. retrieval-augmented generation in LLMs: A comparative evaluation. *Electronics, 14*(24), 4883. https://doi.org/10.3390/electronics14244883

Schooler, J. W., Ohlsson, S., & Brooks, K. (1993). Thoughts beyond words: When language overshadows insight. *Journal of Experimental Psychology: General, 122*(2), 166–183. https://doi.org/10.1037/0096-3445.122.2.166