Co-Overcooked: Cognitive Constraints on Partner Modeling in Human-AI Team Composition

Junghwan Kim[⋆,1] **Dongseok Heo**[⋆,2] **Kyusik Kim**[1] **Jung Lee**[1] **Bongwon Suh**[1]
{jhbale11, ty8900, kyu823, leejungp2, bongwon}@snu.ac.kr
[1]Department of Intelligence and Information, Seoul National University
[2]Interdisciplinary Program in Artificial Intelligence, Seoul National University
Seoul, Republic of Korea
⋆ equal contributors

Abstract

Effective collaboration requires accurate mental models of one's partners. Current AI agents, however, may not model humans the way humans model each other under real-time constraints, creating an asymmetry in human-AI teamwork. Prior research has focused on one-human-one-AI settings, leaving open how coordination changes when multiple humans work with multiple AI agents simultaneously. We developed **Co-Overcooked**, a four-player cooking game where teams of humans and LLM agents must coordinate in real time. In a within-subjects experiment (N=40), participants experienced four team compositions varying in human-to-AI ratio: four humans (H4A0), three humans with one AI (H3A1), two humans with two AIs (H2A2), and one human with three AIs (H1A3). Results revealed nonlinear patterns: H3A1 and H1A3 teams performed worse than pure AI teams, while H2A2 teams showed better performance through spontaneous one-human-one-AI pairing strategies. We identified expectation conflict, where multiple humans hold divergent models of a shared AI partner, as a coordination difficulty specific to hybrid teams, and propose cognitive ownership as a design principle for human-AI team configuration.

Keywords: human-AI collaboration; partner modeling; shared mental models; team composition; cognitive load

Introduction

Effective collaboration requires that individuals maintain accurate mental models of their partners—representations of goals, capabilities, and likely actions that enable anticipatory coordination (Mathieu et al., 2000; Sebanz et al., 2006). In human teams, this partner modeling is supported by Theory of Mind (ToM), enabling bidirectional mentalizing where each agent models their partners while recognizing they themselves are being modeled (Sebanz & Knoblich, 2009; Shteynberg et al., 2023). This recursive awareness contributes to shared mental models that provide a foundation for coordinated action (Cannon-Bowers et al., 1993; Mohammed et al., 2010). Research on joint action shows that individuals automatically co-represent their partner's task, facilitating inference of hidden intentions and coordination without explicit communication (Baker et al., 2011; Sebanz et al., 2013). However, because co-representation competes for limited cognitive resources, the capacity for partner modeling is constrained (Wenke et al., 2011).

When LLM agents join collaborative teams, an asymmetry can emerge: humans naturally form mental models of their partners to predict behavior, while current LLM agents operating under real-time constraints may not engage in reciprocal partner modeling to a comparable degree (Zhang et al., 2024). Recent work suggests that LLMs can exhibit theory-of-mind-like reasoning under certain conditions (Kosinski, 2024; Strachan et al., 2024) and that partner modeling can emerge in simpler multi-agent systems when task demands require it (Mon-Williams et al., 2025); we therefore treat this asymmetry as a property of current systems and prompting regimes rather than a general limitation. This unidirectional mentalizing creates challenges—humans often struggle to form accurate models of opaque AI behaviors (Bansal et al., 2019; Gero et al., 2020), leading to miscalibrated trust and coordination difficulties (Lee & See, 2004; McNeese et al., 2021). While recent approaches have begun endowing AI agents with Theory of Mind capabilities (Li et al., 2025), the burden of modeling non-reciprocal partners often remains with human team members. Existing research has predominantly examined dyadic interactions (Chiang et al., 2023; Gomez et al., 2025), leaving open the question of how partner modeling demands change in multi-agent settings.

The Overcooked environment has become a standard testbed for studying coordination in human-AI teaming. Reinforcement Learning approaches have developed agents capable of zero-shot coordination with novel partners (Hu et al., 2020; Sarkar et al., 2023; Wang et al., 2024), while Large Language Model (LLM) agents exploit semantic reasoning for flexible planning (Agashe et al., 2025; Sun et al., 2025). Computational approaches have modeled Theory of Mind to infer human sub-goals and adapt to diverse playstyles (Nguyen et al., 2024; Schröder & Kopp, 2024; Wu et al., 2021). Additional work has examined role specialization (Mieczkowski et al., 2025) and trust dynamics (Meimandi et al., 2025). Yet the human experience of collaboration—how cognitive strategies and limitations manifest when coordinating with AI—has received limited empirical attention (Rosero et al., 2021), particularly as team composition scales beyond dyads.

Addressing this gap is essential given that practical applications increasingly involve configurations where multiple humans coordinate with multiple LLM agents. Whether the ratio of human to AI team members imposes cognitive constraints on coordination has implications for agent design and team staffing. We suggest that partner modeling capacity constrains viable team configurations in ways that linear scaling assumptions would not predict—configurations requiring humans to

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

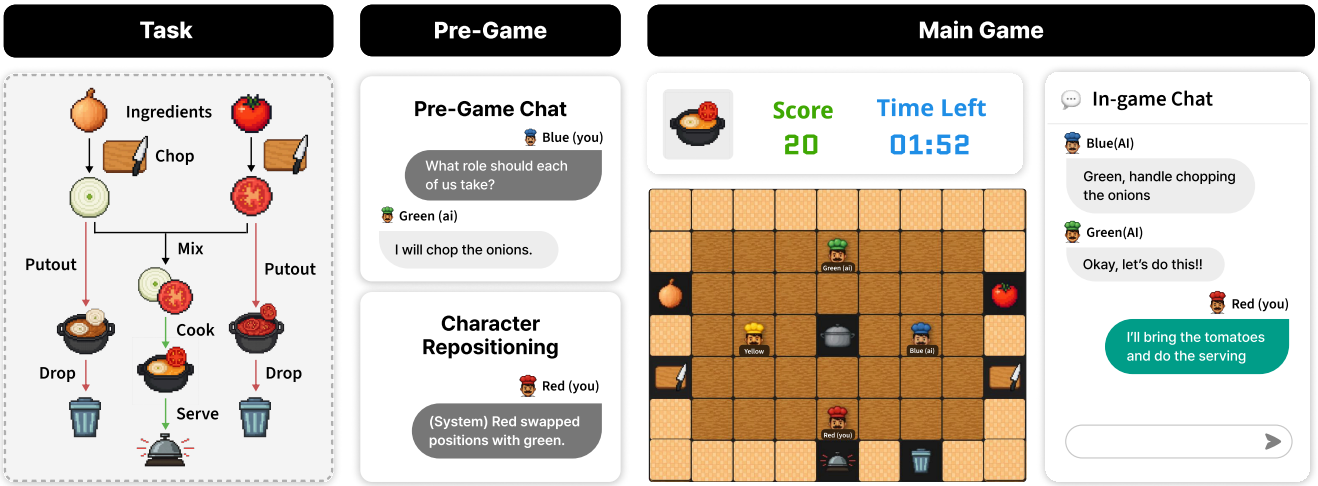

Figure 1: **Co-Overcooked task environment. Left:** Sequential soup preparation workflow. **Center:** Pre-game planning phase with text chat. **Right:** Main game interface with four players coordinating in real time.

share responsibility for AI partners, or to model multiple LLM agents simultaneously, may produce coordination difficulties distinct from those in human-only or dyadic teams.

To investigate this question, we developed **Co-Overcooked**, a four-player collaborative cooking game requiring real-time coordination across interdependent tasks. AI teammates used LLM with heuristic validation, producing behavior that was predictable-in-principle yet fallible-in-practice—instantiating the unidirectional mentalizing characteristic of current human-AI systems. Using a within-subjects design, 40 participants experienced four team compositions (H4A0, H3A1, H2A2, H1A3). Performance in a pure AI baseline condition (H0A4) was collected separately for comparison. We assessed task performance alongside process measures: coded planning communications capturing model formation strategies, behavioral markers of prediction failures, responses to coordination breakdowns, and subjective ratings of mental demand, perceived predictability, and role clarity.

Our findings revealed nonlinear patterns in team performance. Both H3A1 and H1A3 performed worse than pure AI teams—human presence in these configurations appeared to hinder rather than help coordination. H4A0 achieved the highest performance, followed by H2A2. Behavioral analysis pointed to distinct challenges in each configuration: H3A1 teams experienced expectation conflict, where multiple humans held divergent predictions about a shared AI partner without a mechanism for reconciliation. H1A3 teams showed signs of cognitive overload. H2A2 teams performed better through emergent dyadic pairing strategies that kept modeling demands manageable.

These findings offer two contributions. First, we identified expectation conflict as a coordination difficulty specific to human-AI teams. When multiple humans developed divergent models of a shared AI partner, and the AI was not perceived as a participant in resolving these disagreements through mutual coordination, coordination difficulties tended to persist rather than resolve through experience. Second, we propose cognitive ownership as a design principle: effective human–AI teams may benefit from ensuring that each human maintains modeling responsibility for a limited number of AI partners, rather than assuming that human oversight scales additively with more participants.

Experiments

Task Environment

We developed **Co-Overcooked**, a multiplayer adaptation of the Overcooked cooking simulation (Carroll et al., 2019), designed to require continuous partner modeling. Teams of four players worked to prepare onion-tomato soup through a sequential workflow: collecting ingredients, chopping, cooking, and serving (Figure 1, Left). The task environment utilized a spacious map layout allowing unrestricted access to all resources (Figure 1, Right). This open structure meant that coordination could not rely on spatial constraints; instead, players needed to anticipate partners' intentions to avoid physical interference and redundant actions.

The procedure began with a 10-minute tutorial session to familiarize participants with controls and mechanics, followed by the main experiment lasting approximately 20 minutes. Participants completed four sessions corresponding to the different team compositions. Each 5-minute session consisted of three phases: a 1-minute pre-game planning phase where teammates strategized via text chat and could reposition characters (Figure 1, Center), a 3-minute main gameplay phase requiring real-time execution, and a 1-minute post-trial survey to assess subjective workload and role clarity.

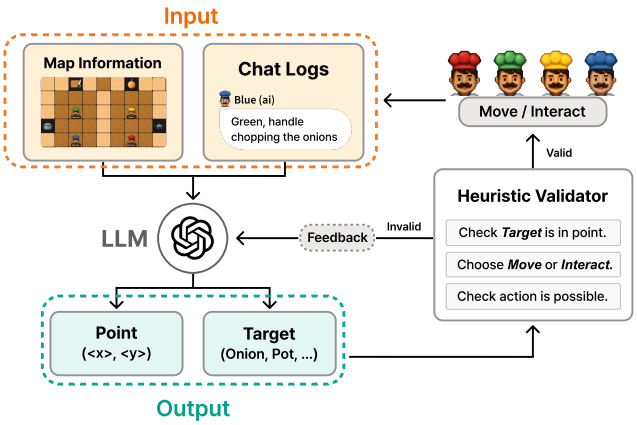

Figure 2: **LLM agent architecture.** The LLM generates targets from map state and chat logs; a heuristic validator ensures executable actions.

LLM Agent Design

LLM agents were governed by a hybrid architecture combining a LLM (GPT-4o) for high-level decision-making with a heuristic validator for robust execution (Figure 2). The agent processes two inputs: the current map state (entity coordinates) and the cumulative chat logs. Based on these inputs, the LLM generates a high-level goal (Target: e.g., "Onion") and specific destination coordinates (Point). To derive the high-level goal, the LLM generates a brief natural language description of its reasoning in a chain-of-thought-like process, which was used internally to guide goal selection. A heuristic validator then assesses the feasibility of this goal; if valid, it converts the intent into low-level controls (Move/Interact). If the action is physically impossible (e.g., path blocked), the validator triggers a feedback loop prompting the LLM to regenerate a valid plan. LLM agents ran in parallel on separate threads, so per-agent inference latency did not block others. The LLM revised intents every 0.5s, while the heuristic validator emitted per-tick Move/Interact actions at the same rate as human controls; humans and AI thus acted at the same tick frequency but differed in intent-revision granularity.

This architecture enabled agents to participate in coordination through natural language. We selected an LLM-based agent architecture to support rich natural language understanding, which we considered necessary for observing how humans assign social roles and coordinate with artificial partners through dialogue. During the pre-game phase, agents analyzed user messages to identify and adopt distinct roles (e.g., recognizing "I'll chop" and responding "I will cook"). During gameplay, they processed real-time chat instructions to dynamically switch current goals. Agent behavior was generated based on the current game state and accumulated chat context, without maintaining explicit, partner-specific models of individual humans over time.

AI-only teams (H0A4) established baseline performance (M=18.2 soups), demonstrating that agents were competent collaborative partners. This baseline ensures that performance differences in mixed teams reflect team structure effects rather than inherent AI limitations.

Participants and Study Design

Forty participants (M_{age} = 24.9 years, SD = 3.85 years; 23 females, 17 males) were recruited via the university's experimental participant pool and organized into 10 teams of four players. An a priori power analysis (Faul et al., 2009) confirmed that this sample size exceeded the minimum of 36 required to detect medium effects (f = .25) with high power ($1-\beta = .95, \alpha = .05$) in a repeated-measures design.

The study was approved by the Institutional Review Board (IRB) of the University of Illustrations. Participants provided written informed consent and received $20 USD compensation. We employed a within-subjects design where each team experienced all four compositions (H4A0, H3A1, H2A2, H1A3). Condition order was counterbalanced using a balanced Latin square design, ensuring each composition appeared equally often in each ordinal position.

Measures

We assessed coordination through multiple complementary measures capturing both objective outcomes and the cognitive processes hypothesized to mediate composition effects.

Task Performance. Completed soups served as the primary outcome, capturing successful execution of the full collaborative workflow.

Partner Model Formation. Pre-game chat messages (1,052 total; κ=.84) were coded into three categories: self-role declaration (stating own roles), partner-role assignment (designating roles to others), and role negotiation (resolving conflicting expectations).

Prediction Failures. Gameplay events (2,847 total; κ=.81) indicating failed partner predictions were coded into four categories: collision events (simultaneous access to the same target object, not physical movement collisions), handoff failures (unmet sequential expectations), idle periods (>2s pauses, corresponding to approximately four missed action steps caused by overlapping agent behaviors), and invalid actions (actions inconsistent with game state).

Shared Mental Model Development. Prediction failure rates were compared between first and second session halves to assess within-session model refinement.

Model Updating Responses. Responses to prediction failures were coded as: corrective instruction (chat commands to modify partner), strategy revision (changing strategy, role of partner), model abandonment (verbal disengagement, e.g., "I can't tell what it's doing"), or no response (absence of verbal reaction, which may include silent adjustments).

Subjective Experience. Post-trial surveys (7-point Likert scales) assessed mental demand (Hart & Staveland, 1988), perceived predictability (Bansal et al., 2019), and role clarity (Seeber et al., 2020).

Analysis Approach

We analyzed coordination through five stages: (1) task performance, (2) pre-game communication patterns, (3) prediction failures and within-session learning, (4) model updating responses, and (5) subjective experience.

Task performance was analyzed with repeated-measures ANOVA at the team level (N=10). Although this sample size is modest, within-subjects design maximizes statistical power by controlling for between-team variability, and the observed effect sizes were large (η_p^2=.81). Subjective ratings were analyzed with repeated-measures ANOVA at the individual level (N=40), treating participants as the unit of analysis as these measures capture individual cognitive experiences. Bonferroni corrections were applied for all pairwise comparisons. Effect sizes are reported as η_p^2. Behavioral coding data descriptively characterize coordination processes.

Results

We present results organized to explain why team composition affects coordination success: coordination outcomes reveal what differs across compositions; behavioral analyses of partner modeling, prediction failures, and model updating reveal how these differences emerge; and subjective ratings confirm the cognitive mechanisms involved.

Coordination Performance

Performance varied nonlinearly with team composition (Figure 3). A repeated-measures ANOVA on team-level performance (N=10 teams) revealed a significant main effect, $F(3,27)$=38.52, p<.001, η_p^2=.81. The observed hierarchy was: H4A0 (M=35.22, SD=7.16) > H2A2 (M=20.82, SD=5.74) > H3A1 (M=11.65, SD=7.22) ≈ H1A3 (M=9.35, SD=8.70).

Pairwise comparisons (Bonferroni-corrected) revealed three findings. First, H3A1 and H1A3 performed equivalently (p=.58) despite having three versus one human member. Second, H2A2 significantly outperformed both H3A1 (p=.002) and H1A3 (p<.001). Third, both H3A1 and H1A3 performed below the AI-only baseline (H0A4: M=18.25), suggesting that human presence does not necessarily improve coordination when team structure limits effective partner modeling.

If coordination were simply additive, performance should increase monotonically with human count. The observed pattern suggests that what matters is not how many humans are present, but whether the team structure supports effective partner modeling. This pattern foreshadows the role of structural alignment in partner modeling, which we unpack through planning communication and prediction failure analyses.

Partner Model Formation During Planning

Pre-game communication (1-minute planning phase) revealed composition-dependent strategies for establishing partner models. We coded 1,052 messages (κ=.84) into three categories.

Each composition showed a distinct communication signature (Table 1). **H4A0** relied on bilateral self-role declaration

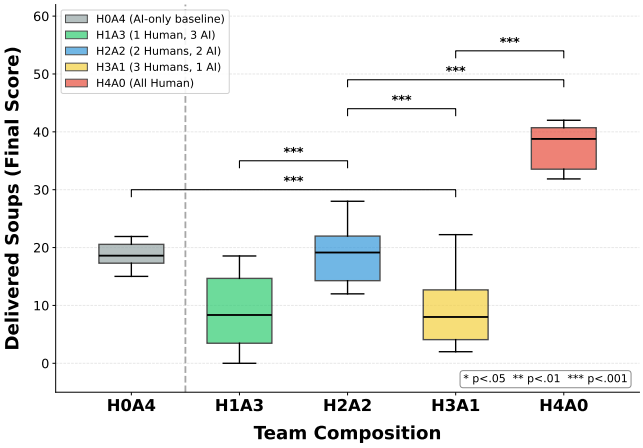

Figure 3: **Coordination performance by team composition.** Dashed line: AI-only baseline (H0A4). Error bars: SE. H3A1 and H1A3 performed below the AI-only baseline.

Table 1: **Pre-game communication patterns by composition.** Mean messages per session (SD below).

Message Type	H4A0	H3A1	H2A2	H1A3
Self-role declaration	**2.89**	1.85	2.10	0.56
	(.95)	(.78)	(.88)	(.42)
Partner-role assignment	3.00	4.20	5.50	**7.25**
	(1.12)	(1.35)	(1.48)	(1.92)
Role negotiation	0.44	**1.80**	0.95	0.85
	(.38)	(.72)	(.55)	(.62)

(M=2.89)—each human stated their intentions, enabling partners to form predictive models through direct exchange. Minimal negotiation (M=0.44) suggests rapid convergence when all partners can communicate.

H3A1 showed elevated role negotiation (M=1.80), the highest among all compositions. This reflects a second-order coordination problem: three humans must align not only their own actions but also their expectations about the shared AI. The AI was not positioned to meaningfully contribute to this negotiation, shifting coordination toward human–human alignment and resulting in persistent divergence in how the AI's role was understood.

H2A2 showed high partner-role assignment (M=5.50) focused on ownership structures. Qualitative analysis revealed that 78% of assignments specified dyadic pairings (e.g., "I'll handle yellow, you take blue"). This strategy manifested as organizing four player coordination into two parallel human–AI pairs.

H1A3 showed minimal self-role declaration (M=0.56) but maximal partner-role assignment (M=7.25). Although the AI offered role-relevant suggestions, the human tended to issue directives and manage the AI's behavior, resulting in an asymmetric information structure that limited reciprocal model formation.

Table 2: **Prediction failures and responses by composition.** Mean events per session (SD below).

	H4A0	H3A1	H2A2	H1A3
Prediction failure types				
Collision	1.20	**3.30**	2.10	2.65
	(.42)	(.94)	(.60)	(.88)
Handoff failure	0.60	**2.20**	1.25	2.00
	(.30)	(.32)	(.44)	(.86)
Idle period	0.40	1.50	1.00	**1.70**
	(.32)	(.42)	(.52)	(.59)
Invalid action	1.30	**3.00**	1.95	2.60
	(.68)	(.67)	(.44)	(.67)
Total failures	3.50	**10.00**	6.30	8.95
	(1.26)	(1.83)	(1.53)	(2.30)
Responses to prediction failures				
Corrective instruction	0.20	**2.40**	1.35	1.55
	(.63)	(.97)	(.74)	(.89)
Strategy revision	0.30	**1.80**	1.15	1.25
	(.50)	(.97)	(1.00)	(.95)
Model abandonment	0.05	**2.00**	0.25	0.55
	(.15)	(.79)	(.22)	(.53)
No response	2.95	3.80	3.55	**5.60**
	(1.42)	(1.30)	(1.21)	(1.39)

Prediction Failures and Shared Model Updating

Gameplay analysis quantified coordination breakdowns from failed partner predictions. We coded 2,847 events (κ=.81) into four categories and examined how teams responded to these breakdowns (Table 2).

Prediction failure types. H3A1 showed the highest total failure rate (M=10.00), with collisions as the dominant mode (M=3.30). In contrast, H1A3's failures were characterized by elevated idle periods (M=1.70), suggesting prediction uncertainty rather than conflicting predictions.

Other failure types reflected mismatches in role execution rather than timing. Handoff failures captured cases in which a player performed an incorrect role, while invalid actions occurred when a player attempted an action that was disallowed by the game rules. Both failure types were most frequent in H3A1 (M=2.20, 3.00) and H1A3 (M=2.00, 2.60), suggesting difficulties in maintaining accurate role assignments in asymmetrical team structures.

Responses to prediction failures. Teams differed in how they responded following prediction failures. H3A1 showed the highest rate of corrective instruction (M=2.40). However, qualitative analysis revealed that 62% of correction sequences involved contradictory instructions from different humans within five-second windows, limiting their effectiveness for stabilizing shared models.

H3A1 also showed the highest model abandonment rate (M=2.00). In contrast, H1A3 exhibited the highest level of no response (M=5.60). This category includes cases in which participants adjusted behavior without verbal coordination; qualitative inspection of gameplay logs showed that such silent behavioral adjustments occurred primarily in the H1A3 condition.

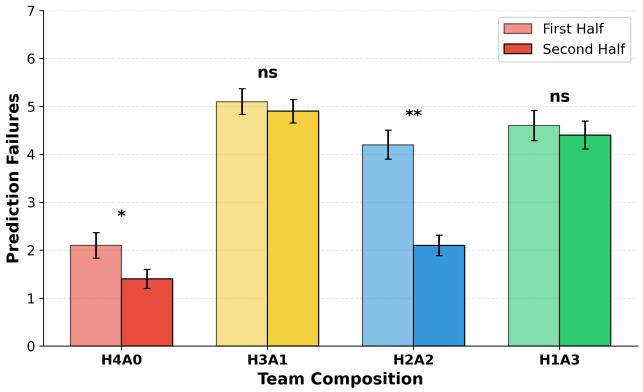

Figure 4: **Within-session learning by team composition.** Only H4A0 and H2A2 showed failure reduction from first to second half. *p<.05, **p<.01, ns = not significant.

Table 3: **Cognitive load and subjective ratings by team composition.** Values represent means on 7-point scales with standard deviations below. Repeated-measures ANOVA: N=40, df=(3, 117).

Measure	H4A0	H3A1	H2A2	H1A3
Mental demand	2.07	3.71	2.38	**4.18**
	(1.21)	(1.81)	(1.71)	(2.01)
		F=14.23, p<.001, η_p^2=.27		
Perceived predictability	**5.82**	3.29	4.62	4.00
	(1.12)	(1.74)	(1.81)	(1.58)
		F=18.76, p<.001, η_p^2=.33		
Role clarity	**5.71**	4.74	5.58	4.29
	(1.63)	(1.73)	(1.38)	(1.76)
		F=7.89, p<.001, η_p^2=.17		

Changes over time. To assess shared model updating, we compared failure rates between the first and second halves of each session (Figure 4). Paired t-tests revealed that only H4A0 and H2A2 showed within-session improvement. H4A0 exhibited a 33% reduction in failures, $t(9)$=2.84, p=.019, d=0.90. H2A2 showed a 50% reduction, $t(9)$=3.52, p=.006, d=1.11. In contrast, neither H3A1, $t(9)$=0.67, p=.52, nor H1A3, $t(9)$=0.89, p=.40, exhibited learning effects, suggesting that their team structures constrained model convergence regardless of experience.

Cognitive Load and Subjective Experience

Post-trial measures confirmed that composition affected experienced difficulty. Repeated-measures ANOVAs revealed significant effects for all three measures.

Three findings illuminate the cognitive mechanisms (Ta-

ble 3). H1A3 reported the highest mental demand (M=4.18), consistent with cognitive overload from tracking three AI partners. Critically, H3A1 reported higher mental demand than H2A2 (M=3.71 vs. 2.38) despite having more human teammates, suggesting that H3A1's difficulty stems from coordinating conflicting human models rather than tracking AI. This interpretation is supported by the predictability ratings: perceived predictability was lowest in H3A1 (M=3.29) despite identical LLM agents across conditions. When three humans hold divergent models of the same AI, each experiences the AI as unpredictable because it repeatedly violates individual expectations. In contrast, H2A2's dyadic ownership structure yielded role clarity (M=5.58) comparable to all-human teams (M=5.71), both exceeding H3A1 and H1A3.

Discussion

This study examines multi-human, multi-AI collaboration beyond dyadic interaction. We asked whether the ratio of humans to AI teammates may shape cognitive constraints on coordination. Our findings suggest that, in this LLM-agent setting, adding humans did not linearly improve performance.

Overall, performance appeared to depend less on the number of humans present than on whether the team structure supported effective partner modeling: H3A1 underperformed H2A2 and even pure-AI teams, challenging the assumption that more human oversight necessarily improves coordination.

Team Composition Shapes Mental Model Evolution

Shared mental models varied by composition. In human-only teams (H4A0), bidirectional exchange allowed participants to declare intentions, observe responses, and refine predictions, producing rapid convergence consistent with prior work (Cannon-Bowers et al., 1993).

Adding AI partners disrupted this process in composition-dependent ways. In H3A1 teams, the core difficulty was coordinating among humans about how to model the AI. Multiple humans formed different expectations of the same AI partner, yet had no mechanism to detect or resolve these differences. We term this expectation conflict. Because the AI could not perceive or resolve human disagreement, conflicting predictions persisted across the session. Elevated contradictory instructions and model abandonment reflected this unresolved divergence, a failure mode specific to hybrid teams with a shared non-reciprocal partner.

H1A3 teams faced a different challenge: cognitive overload. The single human needed to track three AI partners simultaneously. Elevated idle periods and non-responses suggest that this exceeded predictive capacity.

H2A2 teams avoided both problems by forming parallel human-AI dyads. Assigning each human ownership of one AI partner reduced a four-agent problem to two pairs, keeping modeling demands manageable. This suggests ownership allocation as a candidate mechanism for mitigating non-reciprocal coordination.

Together, these patterns suggest a boundary condition for applying shared mental model theory to real-time human-LLM teams: mutual adjustment may break down unless team structure clarifies who is responsible for modeling whom.

Design Implications

Three design considerations follow from these findings. First, team configuration should account for limits in partner modeling capacity. H1A3 shows that too many non-reciprocal partners can overload a human, while H3A1 shows that ambiguous modeling responsibility can produce unresolved expectation conflicts. Human-AI teams should therefore define clear modeling relationships for each human.

Second, modular structures may reduce cognitive load. H2A2 teams decomposed a four-agent problem into two dyads, keeping modeling demands within working memory limits. Rather than treating human-AI teams as undifferentiated pools, systems might organize collaboration as networks of human-AI pairs.

Third, AI agents should support expectation alignment across humans. The expectation conflicts in H3A1 persisted because humans lacked awareness of teammates' divergent models. AI partners that externalize current goals through visual indicators could provide a common reference point, allowing humans to align mental models before conflicts manifest in action.

Limitations and Future Directions

Several limitations constrain generalization. Our task required continuous real-time coordination; slower-paced collaboration might reduce the costs of expectation conflicts. Our LLM agents were moderately predictable; more opaque systems might produce different patterns. We studied ad-hoc teams; extended collaboration might enable teams to develop compensatory protocols.

Future work should test whether these patterns vary by working memory capacity, collaboration duration, and agent type, including non-LLM or RL-based agents with different latency and action granularity. It should also examine whether broadcasting expectation conflicts can restore bidirectional coordination.

Conclusion

This study examined how team composition affects coordination in human-AI teams. Effective collaboration depends on how modeling demands are distributed across team members. Performance was shaped not by the number of humans or LLM agents, but by whether each human could maintain manageable modeling relationships with AI partners. Asymmetric configurations impaired coordination through expectation conflicts (H3A1) or cognitive overload (H1A3). Balanced configurations (H2A2) performed better through spontaneous dyadic pairing that kept modeling demands manageable. As AI agents are integrated into complex team settings, attention to these cognitive constraints may inform configurations that better support coordination.

Acknowledgments

This work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) [NO.RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University)] and Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No. RS-2025-25421701).

References

Agashe, S., Fan, Y., Reyna, A., & Wang, X. E. (2025, April). LLM-coordination: Evaluating and analyzing multi-agent coordination abilities in large language models. In L. Chiruzzo, A. Ritter, & L. Wang (Eds.), *Findings of the association for computational linguistics: Naacl 2025* (pp. 8038–8057). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.findings-naacl.448

Baker, C., Saxe, R., & Tenenbaum, J. (2011). Bayesian theory of mind: Modeling joint belief-desire attribution. *Proceedings of the annual meeting of the cognitive science society*, *33*(33).

Bansal, G., Nushi, B., Kamar, E., Weld, D. S., Lasecki, W. S., & Horvitz, E. (2019). Beyond accuracy: The role of mental models in human-ai team performance. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, *7*, 2–11.

Cannon-Bowers, J. A., Salas, E., & Converse, S. (1993). Shared mental models in expert team decision making. *Individual and group decision making: Current issues*, *221*, 221–46.

Carroll, M., Shah, R., Ho, M. K., Griffiths, T., Seshia, S., Abbeel, P., & Dragan, A. (2019). On the utility of learning about humans for human-ai coordination. *Advances in Neural Information Processing Systems*, *32*.

Chiang, C.-W., Lu, Z., Li, Z., & Yin, M. (2023). Are two heads better than one in ai-assisted decision making? *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–15.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using g* power 3.1: Tests for correlation and regression analyses. *Behavior research methods*, *41*(4), 1149–1160.

Gero, K. I., Ashktorab, Z., Dugan, C., Pan, Q., Johnson, J., Geyer, W., Ruber, M., & Chilton, L. B. (2020). Mental models of ai agents in a cooperative game setting. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–12.

Gomez, C., Cho, S. M., Ke, S., Huang, C.-M., & Unberath, M. (2025). Human-ai collaboration is not very collaborative yet: A taxonomy of interaction patterns. *Frontiers in Computer Science*, *6*, 1521066.

Hart, S. G., & Staveland, L. E. (1988). Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Advances in psychology* (pp. 139–183, Vol. 52). Elsevier.

Hu, H., Lerer, A., Peysakhovich, A., & Foerster, J. (2020). "other-play" for zero-shot coordination. *Proceedings of the 37th International Conference on Machine Learning*, 4399–4410.

Kosinski, M. (2024). Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, *121*(45), e2405460121.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1), 50–80.

Li, X., Ding, Y., Jiang, Y., Zhao, Y., Xie, R., Xu, S., Ni, Y., Yang, Y., & Xu, B. (2025). Dpmt: Dual process multi-scale theory of mind framework for real-time human-ai collaboration. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *47*.

Mathieu, J. E., Heffner, T. S., Goodwin, G. F., Salas, E., & Cannon-Bowers, J. A. (2000). The influence of shared mental models on team process and performance. *Journal of applied psychology*, *85*(2), 273.

McNeese, N. J., Demir, M., Chiou, E. K., & Cooke, N. J. (2021). Trust and team performance in human–autonomy teaming. *International Journal of Electronic Commerce*, *25*(1), 51–72.

Meimandi, K. J., Bolton, M. L., & Beling, P. A. (2025). A trust-aware reinforcement learning approach to enhance human-agent teaming: An overcooked-ai study. *2025 IEEE 5th International Conference on Human-Machine Systems (ICHMS)*, 01–8.

Mieczkowski, E., Mon-Williams, R., Bramley, N. R., Lucas, C. G., VŽlez, N. A., & Griffiths, T. (2025). A normative account of specialization: How task and environment shape role differentiation in collaboration. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *47*.

Mohammed, S., Ferzandi, L., & Hamilton, K. (2010). Metaphor no more: A 15-year review of the team mental model construct. *Journal of management*, *36*(4), 876–910.

Mon-Williams, R., Taylor-Davies, M., Mieczkowski, E., Velez, N., Bramley, N. R., Wang, Y., Griffiths, T. L., & Lucas, C. G. (2025). Partner modelling emerges in recurrent agents (but only when it matters) [Poster]. *Advances in Neural Information Processing Systems*.

Nguyen, D., Le, H., Do, K., Gupta, S., Venkatesh, S., & Tran, T. (2024). Diversifying training pool predictability for zero-shot coordination: A theory of mind approach. *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, 166–174.

Rosero, A., Dinh, F., de Visser, E. J., Shaw, T., & Phillips, E. (2021). Two many cooks: Understanding dynamic human-agent team communication and perception using overcooked 2. *arXiv preprint arXiv:2110.03071*.

Sarkar, B., Shih, A., & Sadigh, D. (2023). Diverse conventions for human-ai collaboration. *Advances in neural information processing systems*, *36*, 23115–23139.

Schröder, F., & Kopp, S. (2024). Towards action-driven mentalizing in realtime human-agent collaboration. *Workshop on Theory of Mind in Human-AI Interaction*.

Sebanz, N., Bekkering, H., & Knoblich, G. (2006). Joint action: Bodies and minds moving together. *Trends in Cognitive Sciences*, *10*(2), 70–76.

Sebanz, N., & Knoblich, G. (2009). Prediction in joint action: What, when, and where. *Topics in Cognitive Science*, *1*(2), 353–367.

Sebanz, N., Knoblich, G., & Prinz, W. (2013). Your task is my task: Shared task representations in dyadic interactions. In *Proceedings of the 25th annual cognitive science society* (pp. 1070–1075). Psychology Press.

Seeber, I., Bittner, E., Briggs, R. O., De Vreede, T., De Vreede, G.-J., Elkins, A., Maier, R., Merz, A. B., Oeste-Reiß, S., Randrup, N., et al. (2020). Machines as teammates: A research agenda on ai in team collaboration. *Information & management*, *57*(2), 103174.

Shteynberg, G., Hirsh, J. B., Wolf, W., Bargh, J. A., Boothby, E. J., Colman, A. M., Echterhoff, G., & Rossignac-Milon, M. (2023). Theory of collective mind. *Trends in Cognitive Sciences*, *27*(11), 1019–1031.

Strachan, J. W., Albergo, D., Borghini, G., Pansardi, O., Scaliti, E., Gupta, S., Saxena, K., Rufo, A., Panzeri, S., Manzi, G., et al. (2024). Testing theory of mind in large language models and humans. *Nature human behaviour*, *8*(7), 1285–1295.

Sun, H., Zhang, S., Niu, L., Ren, L., Xu, H., Fu, H., Zhao, F., Yuan, C., & Wang, X. (2025). Collab-overcooked: Benchmarking and evaluating large language models as collaborative agents. *arXiv preprint arXiv:2502.20073*.

Wang, X., Zhang, S., Zhang, W., Dong, W., Chen, J., Wen, Y., & Zhang, W. (2024). Zsc-eval: An evaluation toolkit and benchmark for multi-agent zero-shot coordination. *Advances in Neural Information Processing Systems*, *37*, 47344–47377.

Wenke, D., Atmaca, S., Holländer, A., Liepelt, R., Baess, P., & Prinz, W. (2011). What is shared in joint action? issues of co-representation, response conflict, and agent identification. *Review of Philosophy and Psychology*, *2*(2), 147–172.

Wu, S. A., Wang, R. E., Evans, J. A., Tenenbaum, J. B., Parkes, D. C., & Kleiman-Weiner, M. (2021). Too many cooks: Bayesian inference for coordinating multi-agent collaboration. *Topics in Cognitive Science*, *13*(2), 414–432.

Zhang, S., Wang, X., Zhang, W., Chen, Y., Gao, L., Wang, D., Zhang, W., Wang, X., & Wen, Y. (2024). Mutual theory of mind in human-ai collaboration: An empirical study with llm-driven ai agents in a real-time shared workspace task. *arXiv preprint arXiv:2409.08811*.