Humans' ALMANAC: A Human Collaboration Dataset of Action-Level Mental Model ANnotations for Agent Collaboration

Jiaju Chen[1], Yuxuan Lu[1], Jiayi Su[1], Chaoran Chen[2], Songlin Xiao[1], Zheng Zhang[3], Yun Wang[4], Yunyao Li [3], Jian Zhao[5], Tongshuang Wu[6], Toby Jia-Jun Li[2], Dakuo Wang[1], Bingsheng Yao[1]* ,

[1]Northeastern University, [2]University of Notre Dame, [3]Adobe, [4]Microsoft Research Asia, [5]University of Waterloo, [6]Carnegie Mellon University

Abstract

Recent advances in LLM agents have enabled complex cognitive capabilities, such as multi-step reasoning, planning, and tool use, that increasingly position these agents as human collaborators. Effective collaboration, however, requires collaborators to continuously maintain and align mental models of their own reasoning, partners' intentions, and shared goals during the collaborative process. Today's agents rarely develop such capabilities since they are primarily optimized for task completion, and the community lacks authentic human collaboration data with action-level mental model annotations that could guide agents toward process-level collaborative competence. To bridge this gap, we present **ALMANAC**[1], a dataset of **A**ction-**L**evel **M**ental model **AN**notations for **A**gent **C**ollaboration built from the Map Task, a classic dyadic routing task from social science. ALMANAC contains 2,987 collaboration actions, each paired with theory-informed mental model annotations that record the participants' self-reasoning, perceived partner intent, and perceived team goal. We benchmark six LLMs on predicting humans' next-turn behavior and mental models. Our results demonstrate ALMANAC's utility in evaluating models' ability to simulate human collaborative behaviors and infer their underlying mental models. The project page is available at https://human-agent-collab.hailab.io/almanac.

1 Introduction

Recent advances in Large Language Model (LLM) agents have enabled complex cognitive capabilities for task-solving (i.e., multi-step reasoning, planning, tool use, and behavioral modeling) that increasingly position these agents as collaborative partners in human workflows (Park et al., 2024; Singh et al., 2025; Chen et al., 2026a; Lu et al., 2025c, 2026a; Yao et al., 2025b). A growing body of work designs LLM agents for complex collaborative tasks such as programming and collaborative writing (He et al., 2025; Venkatraman et al., 2025), where agents engage in multi-turn communication and coordination with humans. In practice, these agents resemble remote human collaborators in important ways, since both operate through struc-

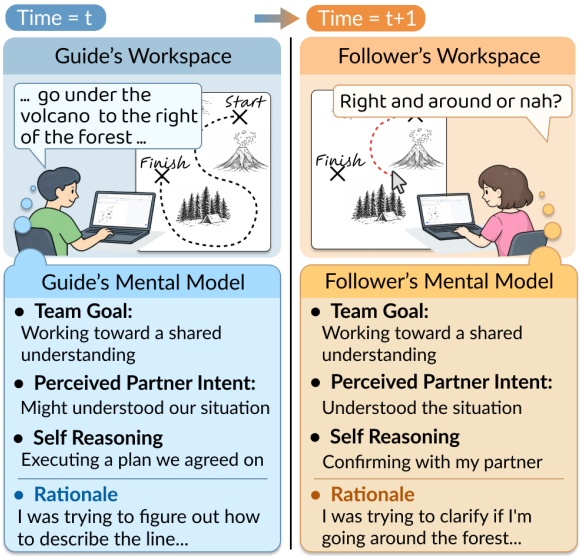

Figure 1: A sample datapoint of ALMANAC, which contains participants' actions, mental models (team goal, perceived partner intent, self-reasoning), and a free-form rationale. We implement the Map Task, a classic dyadic routing task, to collect human collaborative behaviors and action-level mental model annotations.

* Corresponding author. Contact: b.yao@northeastern.edu, chen.jiaju2@northeastern.edu.

[1]ALMANAC is available at https://huggingface.co/datasets/NEU-HAI/Almanac.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 46459–46481
October 24-29, 2026 ©2026 Association for Computational Linguistics

tured, text-based channels and lack the non-verbal cues present in face-to-face interactions (Olson and Olson, 2000; Yao et al., 2025a). Grounding human-agent collaboration in this analogy allows researchers to draw on decades of research on how remote human collaborators build trust, maintain awareness, and coordinate effectively (Clark and Brennan, 1991; Gutwin and Greenberg, 2002), and whether established principles of remote human collaboration persist or break down when partners are LLM agents.

Effective collaboration, however, requires a distinct set of capabilities that task-solving proficiency alone does not provide. Research on human-human collaboration (Cannon-Bowers et al., 1993) has established that successful collaboration depends on collaborators' ability to continuously maintain and align mental models during the collaborative process, including self-reasoning about their own actions, perceived partner intent, and understanding of the shared team goal (Malone and Crowston, 1994; Gutwin and Greenberg, 2002; Marks et al., 2001). The cognitive effort involved in aligning these mental models is what enables collaborators to coordinate actions, establish mutual understanding, and resolve misalignment over time.

However, most current human-agent collaboration remains focused on task-directed exchanges in which the human issues instructions and the agent responds with actions such as tool calls or information retrieval (Qi et al., 2025). LLM agents in such settings are often optimized for independent task completion rather than for maintaining the mental models needed for effective collaboration. Existing agent benchmarks such as ToolBench (Qin et al., 2023), WebArena (Zhou et al., 2023), τ-Bench (Yao et al., 2024), and MultiAgentBench (Zhu et al., 2025) evaluate whether agents can complete tasks under instructions or coordinate task execution, not whether their behaviors support effective collaboration with humans. Thus, agents are rarely exposed to the interaction patterns and cognitive processes that characterize successful collaboration. Existing human collaboration datasets (Lewis et al., 2017; Chawla et al., 2021) reinforce this gap by capturing observable interaction content, such as dialogues and outcomes, while omitting the critical cognitive content that underlies collaborative behaviors (e.g., collaborators' mental models). To our knowledge, no dataset **pairs human collaboration behaviors with action-level mental model annotations grounded in collaboration theory** that can guide agents toward collaborative competence.

In this work, we present **ALMANAC**, a dataset of **A**ction-**L**evel **M**ental model **AN**notations for **A**gent **C**ollaboration built from the Map Task (Anderson et al., 1991), a classic dyadic routing task from social science in which two participants collaborate to reproduce a route through text-based communication and workspace actions such as drawing. We implement the Map Task on a configurable research platform (Yao et al., 2025a) and develop an annotation framework grounded in collaboration theories (Cannon-Bowers et al., 1993; Marks et al., 2001; Gutwin and Greenberg, 2002) to capture participants' mental models at the action level. ALMANAC contains 2,987 collaboration actions from 50 participants across 25 dyadic sessions, each paired with the participant's own mental model annotation, capturing their self-reasoning, perceived partner intent, and perceived team goal, along with a free-form rationale explaining the cognitive process behind the action. We benchmark **six** state-of-the-art LLMs under prompt-based and fine-tuning settings on two complementary tasks: next-turn behavior prediction and mental model prediction. Results show that mental model annotations provide useful signals for predicting human collaborative behavior, but current LLMs remain limited in inferring humans' internal reasoning.

Our contributions are as follows. First, we collect ALMANAC, the first human collaboration dataset that pairs authentic collaborative behaviors with theory-informed, action-level mental model annotations. Second, we design a theory-informed annotation framework that combines in-session checkpoints with post-session retrospective labeling to capture collaborators' action-level mental models and free-form rationales. Third, we benchmark six LLMs and show that mental models offer useful signals for modeling collaborative behavior.

2 Related Work

2.1 Collaboration Datasets and Benchmarks

Existing collaboration-related datasets and benchmarks fall into three categories based on the collaborators involved, as shown in Table 1: human-human, human-agent, and agent benchmarks. The first category consists of datasets grounded in social experiments with humans and mainly designed for dialogue modeling, such as DealNoDeal (Lewis et al., 2017), MutualFriends (He et al., 2017), and CaSiNo (Chawla et al., 2021). The second cate-

Dataset	Interaction Type	Scenario	Mental Model Annotation	Real Human Data
Deal or No Deal	Human–Human	Negotiation	✗	✓
Mutual Friends	Human–Human	Information Sharing	✗	✓
CaSiNo	Human–Human	Negotiation	✗	✓
Rifts	Human–Agent	Dialogue Clarification & Grounding	✗	✓
CoGym	Human–Agent	Multi-Task Collaboration	✗	✓
ToolBench	Single Agent	Tool Use	✗	✗
WebArena	Single Agent	Web Navigation	✗	✗
τ-Bench	Single Agent	Agent-User-Tool Interaction	✗	✗
Multi-Agent-Bench	Multi-Agent	Multi-Task Coordination	✗	✗
ALMANAC	Human–Human	Collaborative Routing	✓	✓

Table 1: Properties of existing representative datasets compared to ALMANAC.

gory focuses on human-agent collaboration, where recent benchmarks measure LLMs' grounding behaviors during human-LLM interaction (Shaikh et al., 2025; Poelitz et al., 2026). Although both involve human participants, they generally lack annotations of collaborators' mental models (Berretta et al., 2023), offering limited support for modeling the underlying reasoning that facilitates effective collaboration.

The third category is agent benchmarks, for instance, ToolBench (Qin et al., 2023), WebArena (Zhou et al., 2023), and τ-Bench (Yao et al., 2024) assess whether agents can follow high-level instructions and execute multi-step actions in interactive tool-use environments. In multi-agent settings, MultiAgentBench (Zhu et al., 2025) evaluates agent performance on tasks such as coding and database error analysis. SOTOPIA (Zhou et al., 2024) provides open-ended agent social scenarios and evaluates agents' social intelligence. However, these benchmarks primarily assess agents' task-solving capabilities and overlook their ability and limitations to coordinate effectively with humans.

2.2 LLM Agents in Human-Agent Collaboration

Recent advances have moved LLM agents beyond static text generation toward interactive task execution, expanding their capabilities in two directions. First, LLM agents can perform increasingly complex tasks that require multi-step planning(Yao et al., 2022), tool invocation(Schick et al., 2023; Yao et al., 2024; Wang et al., 2026; Lu et al., 2026b), and long-horizon interaction (Park et al., 2024; Xu et al., 2025; Chen et al., 2026b). Second, LLM agents have shown emerging cognitive capabilities relevant to collaboration, including language understanding and natural communication (Wang et al., 2024), context perception and situational reasoning (Yao et al., 2022; Chen et al., 2024), and behavior modeling based on provided personas or backstories (Lu et al., 2025c, 2026a; Yao et al., 2025b). To further align agent behaviors with human expectations, recent work has explored supervised fine-tuning on human demonstration data (Xia et al., 2025; Wu et al., 2025) and reinforcement learning from human or environment feedback (Abdulhai et al., 2025; Du et al., 2025; Zhang et al., 2026; Wang et al., 2025b; Lu et al., 2025b).

These advances have motivated the use of LLM agents in various human workflows (Xiao et al., 2024; Sun et al., 2024; Shihab et al., 2025; Chen et al., 2025a; Arakawa et al., 2025; Yang et al., 2026; Sun et al., 2025). However, in many current human-agent collaborations, LLM agents act as assistive systems that respond to human instructions, rather than as equal collaborators that actively infer human partners' intents and evolving mental states throughout the collaboration (Chen et al., 2025b).

3 ALMANAC

We describe the design and construction of ALMANAC in three parts: the annotation framework (Sec. 3.1), the data collection process (Sec. 3.2), and the resulting dataset details (Sec. 3.3).

3.1 Annotation Framework

To collect humans' mental models during collaboration, we design a two-step annotation framework. The *in-session* annotation elicits participants' real-time mental models at key moments in the interaction, while the *post-session* annotation uses the in-session annotation as memory anchors to support action-level mental model annotation.

Grounded in theories of teamwork process (Marks et al., 2001), situation awareness (Endsley, 2017), common ground (Traum, 1995), and workspace awareness (Gutwin and Greenberg,

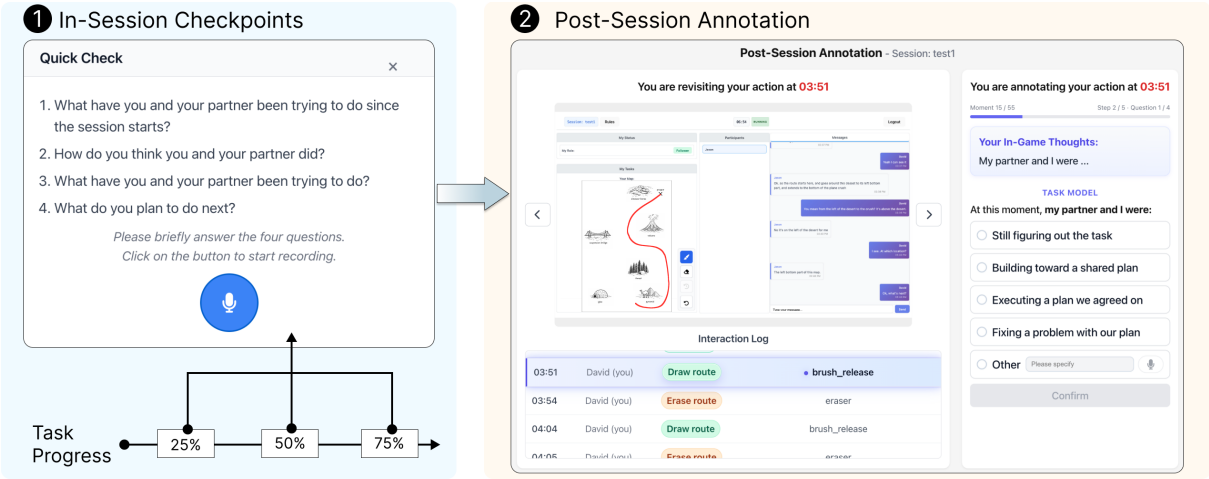

Figure 2: Annotation workflow and interfaces of ALMANAC. Participants first complete the Map Task while providing brief in-session mental model annotations at checkpoints (25%, 50%, 75%). Afterward, they review their action trajectory to retrospectively annotate the team goal, self-reasoning, and perceived partner intent per action.

2002), we translate mental models into three action-level components that are both theoretically central and practically elicitable in the Map Task: the participant's self-reasoning about their own actions, perceived partner intent, and understanding of the shared team goal. In addition to these structured components, participants provide a free-form rationale explaining each action (See Appendix A.1).

Step 1: In-Session Annotation During the Map Task, we periodically elicit participants' mental states through brief in-session checkpoints at 25%, 50%, and 75% of the route drawing progress. We selected three checkpoints spaced at quarter intervals to capture the evolution of mental models across early, middle, and late task stages while keeping interruptions to a minimum (Endsley, 2017; Schinkel-Bielefeld et al., 2024). Specifically, we implement a rule-based mechanism that tracks participants' key actions (e.g., sending a message or drawing a route). At each checkpoint (left part of Figure 2), the system asks participants to briefly report their perceived team goal, partner's intention, and self-reasoning since the last checkpoint. To reduce participant burden, responses are collected via voice recordings and automatically transcribed. Each checkpoint typically takes 10–20 seconds.

Step 2: Post-Session Annotation Immediately after task completion, participants retrospectively annotate their action-level mental models. To support recall, the annotation interface (right part of Figure 2) presents (1) the participant's action trajectory, (2) the screenshot of each action, and (3) the temporally closest in-session response, which serves as a memory anchor (Lyle, 2003) for reconstructing reasoning around that moment. For each action, participants first articulate their action rationale through voice recording. Then, they complete four single-choice questions that capture their mental models (i.e., self-reasoning, perceived partner intent, and perceived team goal, along with an additional item indicating perceived alignment). We additionally validate the retrospective post-session annotation, as reported in Appendix A.2).

3.2 ALMANAC Data Collection

After IRB approval, we recruit a total of 50 participants through snowball sampling (Goodman, 1961). Participants were paired into 25 dyads, each completing one collaboration session. We introduce the data collection and curation process hereinafter.

3.2.1 Participant Background Collection

With participants' consent, we collected their persona information through a structured online survey (see Appendix B). The survey contains two sections: *demographic information*, including age, gender, education level, and *collaboration tendency*. Collaboration tendency is measured using the TeamQ instrument (Britton et al., 2017), a validated scale for capturing participants' attitudes and behavioral tendencies toward collaboration, communication, and collective problem-solving.

3.2.2 Data Collection Process

In the Map Task (Anderson et al., 1991), participants are either assigned as the *Guide* or the *Follower*. The Guide has a map containing both land-

Category		# Dimension / Session	All 25 sessions, 2987 actions				$C_{not_visible}$ 12 sessions, 1469 actions				$C_{visible}$ 13 sessions, 1518 actions			
			Train		Test		Train		Test		Train		Test	
			Avg	SD	Avg	SD	Avg	SD	Avg	SD	Avg	SD	Avg	SD
Action		Total actions	117.7	87.7	125.2	84.5	123.6	96.4	119.0	99.2	112.4	84.0	131.3	88.9
		Message	74.0	57.1	74.3	47.0	88.9	75.0	67.0	45.9	60.6	33.1	81.7	57.1
		Draw	27.4	31.2	25.3	16.2	19.4	11.7	22.7	18.0	34.5	41.3	28.0	17.6
		Erase	8.8	9.0	17.2	21.9	7.9	8.3	24.7	30.6	9.7	9.9	9.7	9.5
		Undo	6.4	7.8	7.8	8.1	6.2	7.7	3.7	5.5	6.5	8.3	12.0	9.0
		Reset	1.1	1.8	0.5	0.5	1.1	1.4	1.0	0.0	1.1	2.2	0.0	0.0
Mental Model	*Team goal*	*Clear on what to do*	43.9	19.7	56.7	22.1	40.2	12.3	56.8	16.2	47.2	24.8	56.6	31.0
		Working toward shared understanding	38.7	18.3	28.9	15.1	42.1	15.6	25.7	11.4	35.6	20.7	32.1	20.1
		Unclear, working it out	11.8	12.0	9.8	8.5	12.2	7.4	11.9	7.2	11.4	15.5	7.7	10.6
		Still figuring out task	4.8	3.9	3.4	3.8	5.0	4.2	3.5	5.7	4.5	3.9	3.3	2.0
	Partner intent	*Understood and aligned*	57.9	24.0	65.4	19.7	49.1	22.9	59.8	21.6	65.7	23.2	70.9	20.3
		Probably understood	19.5	16.9	15.1	14.0	24.2	16.2	19.8	17.1	15.4	17.3	10.4	11.5
		Waiting for more information	16.1	15.6	14.9	6.5	19.2	19.9	16.1	8.3	13.3	10.8	13.7	5.5
		Misunderstood or misaligned	3.7	4.0	3.3	3.9	4.4	3.6	3.0	4.9	3.0	4.4	3.5	3.7
		No clear signal	1.7	3.5	1.3	2.1	1.3	2.7	1.3	2.2	2.1	4.2	1.4	2.5
	Self reasoning	*Executing agreed plan*	41.8	25.2	41.5	11.3	33.9	25.5	35.6	13.2	49.0	23.9	47.4	6.1
		Confirming with partner	27.9	15.9	20.3	7.9	30.3	19.3	21.2	12.4	25.8	12.7	19.3	1.2
		Repairing mistake or misunderstanding	11.9	10.9	12.4	10.1	10.6	8.3	14.4	14.4	13.0	13.1	10.4	5.9
		Grounding by sharing/requesting info	10.4	8.9	14.2	8.7	14.4	8.4	15.3	12.3	6.7	8.1	13.1	5.8
		Exploring independently	3.8	4.9	7.7	5.7	4.6	5.1	11.1	6.5	3.1	4.9	4.4	2.0
		Waiting for more information	2.1	2.8	3.2	3.6	3.5	3.4	2.5	3.5	0.8	1.2	4.0	4.3

Table 2: Dataset statistics and mental model label distributions of ALMANAC across splits and conditions. Mental model values are reported as session-level averages and standard deviations.

marks and a designated route, while the Follower's map contains only the landmarks. The participants need to collaborate to reproduce the route on the follower's map as accurately as possible.

We implement the Map Task following the original protocol (Anderson et al., 1991) on a web-based research platform (Yao et al., 2025a), which enables remote data collection. Participants can communicate via a text-based chat interface. For the Follower, the platform offers a set of tools for drawing routes, including a brush, eraser, undo, and reset buttons. To diversify collaboration behaviors and mental model states, we varied task difficulty by manipulating **whether the Guide could view the Follower's real-time drawing canvas** as a between-subjects factor. In the $C_{visible}$ condition, the Guide's interface displayed both the Guide's map and a live view of the Follower's canvas. In the $C_{not_visible}$ condition, the Guide could only view the Guide's own map. Participant pairs were randomly assigned to one of these two conditions.

Prior to the task, participants were walked through an onboarding procedure covering task rules and platform use to ensure they were familiar with the interface before the session began. In particular, we imposed no time limit on task completion to avoid inducing time-pressure effects that could alter participants' natural collaborative behavior. Sessions lasted an average of 28.25 minutes (SD = 15.59) Participants each received a $25 Amazon gift card as compensation upon completion.

3.2.3 Post-Processing

We remove all personally identifiable information (e.g., names) from the collected data. To enable consistent map representation for LLMs' downstream modeling, we standardize all maps to a discrete grid, aligning spatial elements (e.g., landmarks and routes) with grid coordinates. The Follower's drawing trajectory is converted into cell-level representations by marking the cells traversed by the route. The text-encodable format helps focus the evaluation on collaboration behavior simulation rather than on LLM agents' image comprehension capabilities. The map materials and standardization details are reported in Appendix C.

3.3 Dataset Details

3.3.1 Data Structure

Each session s consists of two participants' personas, action traces, and mental model annotations. Each action at time t denoted as $a_t \in \mathcal{A}$ is timestamped and paired with a post-hoc mental model annotation $m_t = (r_t, g_t, i_t, e_t, \alpha_t)$, where $\mathcal{A} = \{\texttt{message}, \texttt{draw}, \texttt{erase}, \texttt{undo}, \texttt{reset}\}$ denotes the action space. In the mental model tuple, r_t is a free-form rationale , and g_t, i_t, and e_t are text labels capturing *team goal*, *partner intent*, and *self-*

reasoning respectively. α_t denotes human annotated *alignment status*. For each drawing-related action $a_t \in \mathcal{A} \setminus \{\texttt{message}\}$, we record the Follower's canvas state $x_t \in \{0,1\}^{H \times W}$ at time t. Each entry $x_t^{(i,j)}$ indicates whether the corresponding grid cell has been traversed by the route.

3.3.2 Statistics and Analysis of ALMANAC

Dataset Overview. ALMANAC contains 25 sessions (12 in $C_{not_visible}$, 13 in $C_{visible}$), 2,987 human actions with mental model annotations. Table 2 presents the core statistics of ALMANAC. Additional partition and mental model distribution details are reported in Appendix D.

We measure task success as the proportion of cells in the follower's drawing that overlap with the ground-truth route (see Section 4.2). Across all sessions, participants achieved an average final accuracy of 0.66 (SD = 0.12), with $C_{not_visible}$ at 0.67 (SD = 0.12) and $C_{visible}$ at 0.65 (SD = 0.12). Sessions in $C_{not_visible}$ involved more actions on average (122.42, SD = 92.48) than in $C_{visible}$ (116.77, SD = 81.69), reflecting the additional effort required when the Guide could not directly see the Follower's canvas. The large variation in action counts across both conditions suggests substantial differences in teams' collaboration styles.

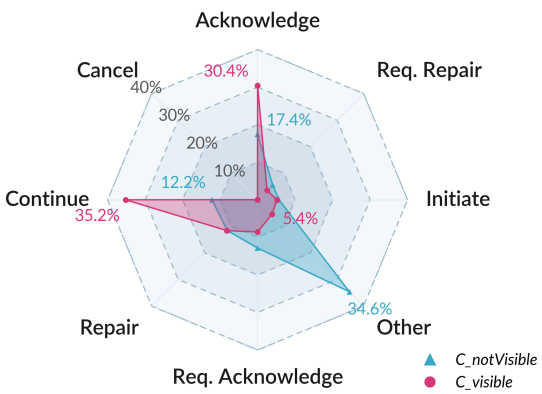

Figure 3: The relative proportion of each grounding act category under $C_{not_visible}$ and $C_{visible}$.

Grounding Act Coding and Validation. To examine how participants establish, confirm, and repair shared understanding during collaboration, we additionally annotate **actions with grounding acts**. Three human annotators independently annotated 180 randomly sampled actions using the grounding act schema proposed by Traum (1995), achieving an averaged inter-coder reliability of 0.81 measured by Fleiss' κ (Fleiss and Cohen, 1973). We randomly partitioned the manually annotated actions into a few-shot set (8 actions) and a held-out validation set (172 actions). The few-shot set was used as in-context examples to prompt GPT-5.5 for automated annotation, while the validation set was reserved exclusively for evaluation. On the held-out validation set, GPT-5.5 achieved a Fleiss' κ of 0.76 against human annotations, suggesting substantial agreement with human labels.

Behavioral and Mental Model Analysis. Figure 3 shows the distribution of grounding acts in $C_{not_visible}$ and $C_{visible}$. In $C_{not_visible}$, participants produced a higher proportion of *Other* (non-grounding) actions. A review of the action logs suggests that drawing-related actions in $C_{not_visible}$ were more likely to serve as individual exploration rather than mutually observable grounding acts. By contrast, the higher proportions of *Continue* and *Acknowledge* acts in $C_{visible}$ suggest that canvas visibility helped participants maintain a clearer shared situation, allowing them to execute the grounded route plan with more explicit confirmation.

Figure 4 further shows how mental model alignment relates to grounding acts and task progress. In the left bar chart, *Acknowledge* acts are associated with higher perceived alignment, whereas repair-related acts, especially *Repair* and *Req. Repair*, are associated with higher perceived misalignment. In the right line chart, alignment in team goal, partner intent, and self-reasoning generally increased as the task unfolded, with consistently higher alignment in $C_{visible}$ than in $C_{not_visible}$. These trends suggest that participants' mental models became more aligned over time, and that access to the shared drawing state may have supported mutual understanding. The systematic variation in mental model alignment across conditions and task stages provides a natural basis for evaluating whether LLMs can capture these collaborative dynamics in Sec.4.

4 Benchmark Experiment

We evaluate how ALMANAC can be leveraged to assess LLMs' ability to simulate humans' collaboration behavior and mental models through two complementary tasks:

1. Next Behavior Prediction. For a target participant, given the behavior trajectory history and persona profile, predict the next behavior.

2. Mental Model Prediction. For a target participant, given the behavior trajectory history, mental model history, and persona profile, predict the participant's mental state in the next turn.

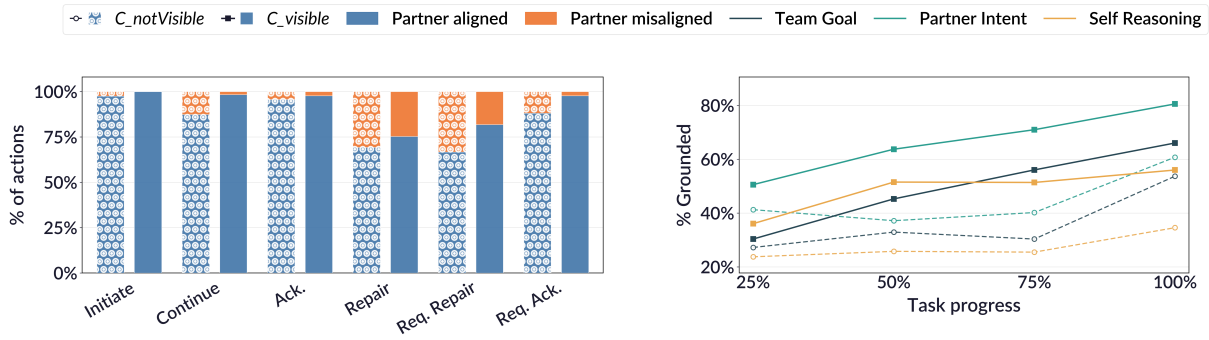

Figure 4: Relationships between grounding acts and mental model alignment. Left: the proportion of perceived partner intent alignment within each grounding act category across $C_{not_visible}$ and $C_{visible}$. Right: the proportion of grounded team goal, partner intent, and self-reasoning annotations over task progress in the two conditions.

Next action prediction evaluates whether models can predict a human collaborator's next turn move from the preceding interaction context, reflecting their ability to simulate observable collaboration dynamics (Lu et al., 2025a; Wang et al., 2025a). Mental model prediction goes further by assessing whether models can infer the collaborator's next step underlying reasoning, which helps determine whether a model shows a genuine understanding of the collaboration state or merely fits surface-level trajectory patterns.

4.1 Experiment Setup

Our benchmark experiment includes two open-sourced models (Qwen3.6-35B-A3B and Llama 3.3 70B), two proprietary models (GPT 5.5 and Claude 4.6 Sonnet), and two finetuned models (Qwen3-4B and Qwen3-30B-A3B) on ALMANAC. We evaluate these models under two approaches:

Persona-Based LLM. In this setting, we evaluate whether general-purpose LLMs can simulate human collaborative behaviors when provided with only the participant's profile. Each model is given the participant persona, including demographic information and collaboration profiles, together with the interaction history up to the current action. We apply this setting to Qwen3.6-35B-A3B, Llama 3.3 70B, GPT-5.5, and Claude 4.6 Sonnet. The full prompts are provided in Appendix G.

Fine-Tuned LLM. In this setting, we examine whether smaller models can benefit from supervision on ALMANAC. We fine-tune Qwen3-4B and Qwen3-30B-A3B on the training split and evaluate their performance on the same next action prediction and mental model prediction tasks. Hyperparameters are reported in Appendix E

For the next action prediction task, we compare three settings to explore the effectiveness of human-annotated mental models. In the default setting, the model predicts the next action from only the interaction history. In the second, the model is given the participant's annotated mental model before the target action, denoted as *{+Mental Model}*. Including ground-truth mental model annotations evaluates whether explicit human-annotated mental model information can improve model performance. We also consider a Chain of Thought setting, denoted as *{+CoT}*, in which the model first generates a rationale from the interaction history and then predicts the next action conditioned on that rationale. Finally, we experiment a two-stage ReAct-inspired baseline, denoted as *+Pred. MM*, in which the model first predicts humans' mental model from the observed interaction history and then predicts the next action conditioned on the predicted mental model. The full results are shown in Appendix F.

4.2 Evaluation

We evaluate model outputs at two levels of granularity. At the category level, we report **accuracy** and **recall** for the predicted next action type and mental model category. At the content level, we evaluate semantic similarity using **SBERT** (Reimers and Gurevych, 2019). For message actions, we compare the generated message to the ground-truth message; for mental model prediction, we compare the generated rationale to the participant's action-level annotation. Full results including **ROUGE-L** (Lin, 2004) are presented in Appendix F.

To evaluate models' drawing trace accuracy, we use a distance-weighted score. Each predicted ink cell is scored by its Chebyshev distance to the ground-truth route: 1 if on the route, $2/3$ at dis-

Model	$Accuracy_{\text{Action_Type}}$				$Recall_{\text{Action_Type}}$				$SBERT_{\text{Message}}$				$Accuracy_{\text{Drawing}}$	
	$C_{visible}$		$C_{not_visible}$		$C_{visible}$		$C_{not_visible}$		$C_{visible}$		$C_{not_visible}$		$C_{visible}$	$C_{not_visible}$
	G	F	G	F	G	F	G	F	G	F	G	F	F	F
Qwen3-35B-A3B	1.00	0.46	1.00	0.54	1.00	0.34	1.00	0.30	0.22	0.25	0.28	0.27	0.43	0.45
+*Mental Model*	1.00	0.48 ⇑	1.00	0.55 ⇑	1.00	0.38 ⇑	1.00 ⇑	0.31 ⇑	0.21	0.27 ⇑	0.28	0.29 ⇑	0.44 ⇑	0.43
Llama 3.3 70B	1.00	0.44	1.00	0.51	1.00	0.32	1.00	0.28	0.22	0.23	0.31	0.27	0.57	0.45
+*Mental Model*	1.00	0.51 ⇑	1.00	0.51	1.00	0.28	1.00	0.30 ⇑	0.22	0.25 ⇑	0.30	0.29 ⇑	0.46	**0.56** ⇑
GPT-5.5	1.00	0.56	1.00	0.59	1.00	0.44	1.00	0.35	0.24	0.32	0.33	0.36	**0.55**	0.43
+*Mental Model*	1.00	**0.58** ⇑	1.00	**0.61** ⇑	1.00	**0.46** ⇑	1.00	**0.37** ⇑	**0.25** ⇑	0.17	0.33	**0.38** ⇑	**0.55**	0.47 ⇑
Claude 4.6 Sonnet	1.00	0.47	1.00	0.54	1.00	0.36	1.00	0.30	0.23	0.29	0.31	0.31	0.45	0.44
+*Mental Model*	1.00	0.51 ⇑	1.00	0.55 ⇑	1.00	0.39 ⇑	1.00	0.31 ⇑	0.23	0.31 ⇑	0.31	0.33 ⇑	0.43	0.53 ⇑
Qwen3-4B FT	1.00	0.56	1.00	0.54	1.00	0.37	1.00	0.31	0.21	0.35	0.23	0.23	0.47	0.44
Qwen3-30B-A3B FT	1.00	0.52	1.00	0.52	1.00	0.30	1.00	0.27	0.20	**0.37**	0.22	0.26	0.54	0.55

Table 3: Action Type Accuracy, Action Type Recall, Message SBERT, and Follower's Drawing Accuracy across six models in Guide (G) and Follower (F) roles under $C_{visible}$ and $C_{not_visible}$. **Bolded** numbers (excluding Guide's Action Type Accuracy and Recall) indicate the best performance for each role and canvas visibility condition.

tance 1, $1/3$ at distance 2, and 0 otherwise. We report the average score over all predicted ink cells.

4.3 Result and Analysis

4.3.1 Next Action Prediction

Table 3 shows model performance on the next action prediction task. Across all models, Guide action type prediction is perfect, as expected, since the Guide's action space is limited to `message`. Follower action prediction, by contrast, is substantially harder, which is consistent with Followers alternating between interpreting messages, drawing, correcting, and grounding their understanding.

Models generally perform better in $C_{not_visible}$ than in $C_{visible}$ across action type, message, and drawing prediction. For example, GPT-5.5 achieves higher Follower $Accuracy_{\text{Action_Type}}$ in $C_{not_visible}$, and its Follower's $SBERT_{\text{Message}}$ rises from 0.17 to 0.38 with mental model input. For drawing, GPT-5.5 reaches around 0.55 in $C_{visible}$, while in $C_{not_visible}$, mental model input improves $Accuracy_{\text{Drawing}}$ from 0.43 to 0.47. The greater behavioral variability in $C_{visible}$ may explain this pattern, since Guides who can observe the Follower's live canvas tend to rely more on visually grounded corrections, interruptions, and fine-grained coordination, making the next action harder to infer from textual history alone.

Adding mental model input generally improves model performance in Followers' action prediction. For GPT-5.5, mental model input slightly improves Follower action type accuracy and recall in both conditions, and improves drawing accuracy in $C_{not_visible}$. This result suggests that Followers' behavior is more grounded in latent reasoning states (e.g., interpreting instructions) than Guides', whose actions are more anchored to the visible canvas and task goal. Notably, smaller models fine-tuned on ALMANAC come close to large proprietary models, indicating that targeted supervision on ALMANAC can effectively close the gap with larger models.

4.3.2 Mental Model Prediction

Table 4 shows model performance on the mental model prediction task. Follower mental models are easier to predict than Guides' across all three dimensions. For example, Claude 4.6 Sonnet achieves 0.75 $Accuracy_{\text{Team_Goal}}$ for the Follower under $C_{visible}$, but only 0.48 for the Guide, with similar gaps for partner intent and self-reasoning. Although the Guide's action space is limited to `message`, the Guide's underlying reasoning likely involves richer spatial planning and partner monitoring that are hard to infer from interaction history alone, which may account for the asymmetry.

Across roles and conditions, self-reasoning is the hardest dimension to predict, whereas team goal and partner intent are more predictable. Team goal and partner intent are often reflected in shared task progress and dialogue content, while self-reasoning captures participant-specific motivations that may not be explicitly expressed. As a result, current LLMs appear better at approximating shared components of mental model awareness than inferring private reasoning that varies across participants.

Notably, fine-tuned Qwen-3-4B achieves the strongest performance in Followers' mental model prediction and $SBERT_{\text{Rationale}}$, whereas prompt-based models' Follower $SBERT_{\text{Rationale}}$ scores stay within a narrow range of 0.40–0.55. This pattern suggests that ALMANAC 's mental model annotations provide useful supervision for learning collaboration-relevant reasoning when participant states are reflected in the interaction history. Overall, no single model consistently performs best

Model	$Accuracy_{\text{Team_Goal}}$				$Accuracy_{\text{Partner_Intent}}$				$Accuracy_{\text{Self_Reasoning}}$				$SBERT_{\text{Rationale}}$			
	$C_{visible}$		$C_{not_visible}$		$C_{visible}$		$C_{not_visible}$		$C_{visible}$		$C_{not_visible}$		$C_{visible}$		$C_{not_visible}$	
	G	F	G	F	G	F	G	F	G	F	G	F	G	F	G	F
Qwen3-35B-A3B	**0.48**	0.69	0.41	0.64	**0.58**	0.73	0.38	0.67	**0.31**	0.56	0.32	0.52	**0.41**	0.46	0.44	0.52
Llama 3.3 70B	0.43	0.71	**0.56**	0.64	0.38	0.76	0.34	0.72	0.23	0.55	0.31	0.53	**0.41**	0.47	0.44	0.54
GPT-5.5	0.41	0.72	0.35	0.68	0.37	0.75	0.36	0.70	0.29	0.60	0.32	0.59	0.40	0.51	**0.45**	0.55
Claude 4.6 Sonnet	**0.48**	0.75	0.45	0.68	0.51	0.76	0.45	0.71	0.27	0.56	0.30	0.55	**0.41**	0.52	**0.45**	0.55
Qwen3-4B FT	0.37	**0.81**	0.51	**0.88**	0.40	**0.84**	**0.47**	**0.84**	0.28	**0.65**	0.30	**0.70**	0.37	**0.76**	0.33	0.64
Qwen3-30B-A3B FT	0.47	0.55	0.39	0.55	0.38	0.78	0.46	0.77	0.29	0.54	**0.33**	0.54	0.34	0.61	0.38	**0.66**

Table 4: Team Goal Accuracy, Partner Intent Accuracy, Self-Reasoning Accuracy, and Rationale SBERT across six models in Guide (G) and Follower (F) roles under $C_{visible}$ and $C_{not_visible}$. **Bolded** numbers indicate the best performance for each role and canvas visibility condition.

across all conditions, and the low self-reasoning accuracy highlights private mental model inference as a central challenge in ALMANAC.

5 Discussion

Our results show that mental model annotations provide useful signals for modeling collaborative behavior. In next action prediction, adding these annotations improves some models' performance, but the gains are inconsistent, indicating that mental models encode signals that current LLMs do not reliably leverage. The mental model prediction results reinforce this interpretation, given that the shared components (e.g., team goal and partner intent) are easier to infer than self-reasoning, suggesting that models handle publicly grounded collaboration states better than private reasoning.

The two experiments suggest a role-specific dissociation between behavior prediction and mental model prediction. Guide mental models are harder to infer because they involve less observable reasoning about route planning and Follower progress. Followers show the opposite pattern: their broader action space makes behavior prediction harder, but their mental models are more directly shaped by the Guide's explicit instructions. This result suggests that observable behavior and mental models provide complementary signals, so success on one does not necessarily imply success on the other.

Importantly, next action prediction is not the end goal of ALMANAC; it serves as a diagnostic for whether models can simulate the observable layer of collaboration. The deeper challenge lies in building agents that maintain accurate mental models throughout the collaboration. ALMANAC provides a foundation for developing such agents by supplying the process-level supervision signals that current training paradigms lack.

6 Conclusion

In this work, we present ALMANAC, an authentic human collaboration dataset that captures both human collaboration behaviors and the underlying action-level mental models, including how humans reason about their team goals, partners' intentions, and self-reasoning over time. We demonstrate the utility of ALMANAC through next action prediction and mental model prediction. Our results show that mental model annotations provide signals beyond interaction history alone, and that shared mental model components are substantially easier to predict than private self-reasoning. This observation highlights a fundamental gap in models' ability to infer the cognitive processes that drive effective collaboration. By grounding agent evaluation in authentic human collaboration data with theory-informed mental model annotations, ALMANAC opens a pathway toward developing LLM agents that can serve as genuine collaborative partners rather than sophisticated task-solvers.

Acknowledgment

This work was supported in part by a Microsoft Research Agentic AI Research and Innovation Award. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsors.

7 Limitations

This work has several limitations, which we discuss alongside the design choices that mitigate them.

First, our annotation framework relies in part on post-session retrospective reports, which are susceptible to recall bias and post-hoc rationalization. We mitigate this concern through two design choices: the *in-session* checkpoints capture real-time mental states at three task stages and serve

as memory anchors during *post-session* annotation, and the annotation interface presents action-level screenshots alongside the interaction trajectory to support context-specific recall. Future work could explore concurrent think-aloud protocols or physiological measures to further validate the fidelity of retrospective annotations.

Second, the ALMANAC dataset comprises 25 sessions from 50 participants, which is modest compared to some large-scale NLP benchmarks. Nevertheless, the detailed action-level annotation of theory-grounded mental models and rationales partially compensates for the session count, which yields 2,987 individually annotated data points with both structured labels and free-form rationales. In addition, our participants include both native and non-native English speakers, and we do not control for proficiency level in the current analysis.

Third, ALMANAC is built from the Map Task, a single controlled task domain selected for its theoretical grounding in social science research and its natural combination of language and workspace actions (Anderson et al., 1991). While the controlled setting allows us to isolate collaboration variables, real-world collaboration often involves longer time horizons and more complex interaction constraints and social dynamics. Extending the annotation framework to other collaborative tasks, such as collaborative writing, programming, or decision-making, would strengthen claims about the generalizability of both the dataset and the benchmark findings. A promising direction for future work is to study how models can adaptively construct and update mental models in diverse, real-world, domain-specific environments.

Fourth, our benchmark evaluates six LLMs under persona-based prompting and supervised fine-tuning on ALMANAC, but does not include models fine-tuned on other collaborative dialogue datasets (e.g., CaSiNo (Chawla et al., 2021), DealNoDeal (Lewis et al., 2017)) or models trained with alternative alignment approaches such as reinforcement learning from human feedback. Including such baselines would help disentangle whether performance gaps stem from the absence of collaboration-specific training signals or from architectural limitations of current models. In addition, current language models remain limited in interpreting drawing actions and map states. Although we represent maps and drawing trajectories in structured text-based formats, these representations may not fully capture the spatial relationships that human participants perceive visually. Future work could explore multimodal models that jointly process visual and textual input to better represent the spatial task state.

References

Marwa Abdulhai, Ryan Cheng, Donovan Clay, Tim Althoff, Sergey Levine, and Natasha Jaques. 2025. Consistently simulating human personas with multi-turn reinforcement learning. *arXiv preprint arXiv:2511.00222*.

Anne H Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, and 1 others. 1991. The hcrc map task corpus. *Language and speech*, 34(4):351–366.

Robert W Andrews, J Mason Lilly, Divya Srivastava, and Karen M Feigh. 2023. The role of shared mental models in human-ai teams: a theoretical review. *Theoretical Issues in Ergonomics Science*, 24(2):129–175.

Riku Arakawa, Hiromu Yakura, Kei Akuzawa, and Shizuma Kubo. 2025. Ai for meeting minutes: Promises and challenges in designing human-ai collaboration on a production saas platform. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '25, New York, NY, USA. Association for Computing Machinery.

Sophie Berretta, Alina Tausch, Greta Ontrup, Björn Gilles, Corinna Peifer, and Annette Kluge. 2023. Defining human-ai teaming the human-centered way: a scoping review and network analysis. *Frontiers in Artificial Intelligence*, 6:1250725.

Emily Britton, Natalie Simper, Andrew Leger, and Jenn Stephenson. 2017. Assessing teamwork in undergraduate education: a measurement tool to evaluate individual teamwork skills. *Assessment & Evaluation in Higher Education*, 42(3):378–397.

Janis A Cannon-Bowers, Eduardo Salas, and Sharolyn Converse. 1993. Shared mental models in expert team decision making. *Individual and group decision making: Current issues*, 221:221–46.

Kushal Chawla, Jaysa Ramirez, Rene Clever, Gale Lucas, Jonathan May, and Jonathan Gratch. 2021. CaSiNo: A corpus of campsite negotiation dialogues for automatic negotiation systems. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3167–3185, Online. Association for Computational Linguistics.

Jiaju Chen, Yuxuan Lu, Xiaojie Wang, Huimin Zeng, Jing Huang, Jiri Gesi, Ying Xu, Bingsheng Yao, and Dakuo Wang. 2026a. Multi-agent-as-judge: Aligning LLM-agent-based automated evaluation with

multi-dimensional human evaluation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17389–17416, San Diego, California, United States. Association for Computational Linguistics.

Jiaju Chen, Yuxuan Lu, Shao Zhang, Bingsheng Yao, Yuanzhe Dong, Ying Xu, Yunyao Li, Qianwen Wang, Dakuo Wang, and Yuling Sun. 2024. StorySparkQA: Expert-annotated QA pairs with real-world knowledge for children's story-based learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17351–17370, Miami, Florida, USA. Association for Computational Linguistics.

Jiaju Chen, Bo Sun, Yuxuan Lu, Yun Wang, Dakuo Wang, and Bingsheng Yao. 2026b. Collabsim: A cscw-grounded methodology for investigating collaborative competence of llm agents through controlled multi-agent experiments. *arXiv preprint arXiv:2606.06399*.

Jiaju Chen, Minglong Tang, Yuxuan Lu, Bingsheng Yao, Elissa Fan, Xiaojuan Ma, Ying Xu, Dakuo Wang, Yuling Sun, and Liang He. 2025a. Characterizing llm-empowered personalized story reading and interaction for children: Insights from multi-stakeholder perspectives. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.

Valerie Chen, Alan Zhu, Sebastian Zhao, Hussein Mozannar, David Sontag, and Ameet Talwalkar. 2025b. Need help? designing proactive ai assistants for programming. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–18.

Herbert H Clark and Susan E Brennan. 1991. Grounding in communication.

Philip R Doyle, Justin Edwards, Odile Dumbleton, Leigh Clark, and Benjamin R Cowan. 2019. Mapping perceptions of humanness in intelligent personal assistant interaction. In *Proceedings of the 21st international conference on human-computer interaction with mobile devices and services*, pages 1–12.

Bangde Du, Ziyi Ye, Zhijing Wu, Monika A. Jankowska, Shuqi Zhu, Qingyao Ai, Yujia Zhou, and Yiqun Liu. 2025. SimVBG: Simulating individual values by backstory generation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 13093–13122, Suzhou, China. Association for Computational Linguistics.

Mica R Endsley. 2017. Direct measurement of situation awareness: Validity and use of sagat. In *Situational awareness*, pages 129–156. Routledge.

Joseph L Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*, 33(3):613–619.

Leo A Goodman. 1961. Snowball sampling. *The annals of mathematical statistics*, pages 148–170.

Carl Gutwin and Saul Greenberg. 2002. A descriptive framework of workspace awareness for real-time groupware. *Computer Supported Cooperative Work (CSCW)*, 11(3):411–446.

He He, Anusha Balakrishnan, Mihail Eric, and Percy Liang. 2017. Learning symmetric collaborative dialogue agents with dynamic knowledge graph embeddings. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1766–1776, Vancouver, Canada. Association for Computational Linguistics.

Junda He, Christoph Treude, and David Lo. 2025. Llm-based multi-agent systems for software engineering: Literature review, vision, and the road ahead. *ACM Transactions on Software Engineering and Methodology*, 34(5):1–30.

Gary Klein, Paul J. Feltovich, Jeffrey Bradshaw, and David Woods. 2005. *Common Ground and Coordination in Joint Activity*, pages 139 – 184.

Mike Lewis, Denis Yarats, Yann Dauphin, Devi Parikh, and Dhruv Batra. 2017. Deal or no deal? end-to-end learning of negotiation dialogues. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2443–2453, Copenhagen, Denmark. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yuxuan Lu, Ting-Yao Hsu, Hansu Gu, Limeng Cui, Yaochen Xie, William P. Headden, Bingsheng Yao, Akash Veeragouni, Jiapeng Liu, Sreyashi Nag, Jessie Wang, and Dakuo Wang. 2026a. Agent A/B: Automated and Scalable A/B Testing on Live Websites with Interactive LLM Agents. In *Proceedings of the Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems*, pages 1–12, Barcelona , Spain. ACM.

Yuxuan Lu, Jing Huang, Yan Han, Bingsheng Yao, Sisong Bei, Jiri Gesi, Yaochen Xie, Yisi Sang, Qi He, Dakuo Wang, and 1 others. 2025a. Can llm agents simulate multi-turn human behavior? evidence from real online customer behavior data. *arXiv preprint arXiv:2503.20749*.

Yuxuan Lu, Jing Huang, Hui Liu, Jiri Gesi, Yan Han, Shihan Fu, Tianqi Zheng, and Dakuo Wang. 2025b. WEBSERV: A Browser-Server Environment for Efficient Training of Reinforcement Learning-based Web Agents at Scale. *Preprint*, arXiv:2510.16252.

Yuxuan Lu, Ziyi Wang, Yingzhou Lu, Yisi Sang, Jiri Gesi, Xianfeng Tang, Yimeng Zhang, Zhenwei Dai, Hui Liu, Hanqing Lu, Chen Luo, Qi He, Benoit Dumoulin, Jing Huang, and Dakuo Wang. 2026b. Firefly: Illuminating Large-Scale Verified

Tool-Call Data Generation from Real APIs. *Preprint*, arXiv:2605.17558.

Yuxuan Lu, Bingsheng Yao, Hansu Gu, Jing Huang, Jessie Wang, Yang Li, Jiri Gesi, Qi He, Toby Jia-Jun Li, and Dakuo Wang. 2025c. UXAgent: A System for Simulating Usability Testing of Web Design with LLM Agents. *arXiv preprint*. ArXiv:2504.09407 [cs.CL].

John Lyle. 2003. Stimulated recall: A report on its use in naturalistic research. *British Educational Research Journal - BR EDUC RES J*, 29:861–878.

Thomas W Malone and Kevin Crowston. 1994. The interdisciplinary study of coordination. *ACM Computing Surveys (CSUR)*, 26(1):87–119.

Michelle A Marks, John E Mathieu, and Stephen J Zaccaro. 2001. A temporally based framework and taxonomy of team processes. *Academy of management review*, 26(3):356–376.

Gary M Olson and Judith S Olson. 2000. Distance matters. *Human–computer interaction*, 15(2-3):139–178.

Joon Sung Park, Carolyn Q Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S Bernstein. 2024. Generative agent simulations of 1,000 people. *arXiv preprint arXiv:2411.10109*.

Christian Poelitz, Finale Doshi-Velez, and Siân Lindley. 2026. A benchmark to assess common ground in human-ai collaboration. *arXiv preprint arXiv:2602.21337*.

Yunjia Qi, Hao Peng, Xiaozhi Wang, Amy Xin, Youfeng Liu, Bin Xu, Lei Hou, and Juanzi Li. 2025. Agentif: Benchmarking instruction following of large language models in agentic scenarios. *arXiv preprint arXiv:2505.16944*.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, and 1 others. 2023. Toolllm: Facilitating large language models to master 16000+ real-world apis. *arXiv preprint arXiv:2307.16789*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 3982–3992.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *Advances in neural information processing systems*, 36:68539–68551.

Nadja Schinkel-Bielefeld, Louise Burke, Inga Holube, Maria Iankilevitch, Lorienne M Jenstad, Dina Lelic, Graham Naylor, Gurjit Singh, Karolina Smeds, Petra von Gablenz, and 1 others. 2024. Implementing ecological momentary assessment in audiological research: Opportunities and challenges. *American journal of audiology*, 33(3):648–673.

Kjeld Schmidt and Liam Bannon. 1992. Taking cscw seriously: Supporting articulation work. *Computer supported cooperative work (CSCW)*, 1(1):7–40.

Omar Shaikh, Hussein Mozannar, Gagan Bansal, Adam Fourney, and Eric Horvitz. 2025. Navigating rifts in human-LLM grounding: Study and benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20832–20847, Vienna, Austria. Association for Computational Linguistics.

Md Istiak Hossain Shihab, Christopher Hundhausen, Ahsun Tariq, Summit Haque, Yunhan Qiao, and Brian Wise Mulanda. 2025. The effects of github copilot on computing students' programming effectiveness, efficiency, and processes in brownfield coding tasks. In *Proceedings of the 2025 ACM Conference on International Computing Education Research V. 1*, pages 407–420.

Joykirat Singh, Raghav Magazine, Yash Pandya, and Akshay Uttama Nambi. 2025. Agentic reasoning and tool integration for llms via reinforcement learning. *ArXiv*, abs/2505.01441.

Lu Sun, Shihan Fu, Bingsheng Yao, Yuxuan Lu, Wenbo Li, Hansu Gu, Jiri Gesi, Jing Huang, Chen Luo, and Dakuo Wang. 2025. LLM Agent Meets Agentic AI: Can LLM Agents Simulate Customers to Evaluate Agentic-AI-based Shopping Assistants? *Preprint*, arXiv:2509.21501.

Yuling Sun, Jiaju Chen, Bingsheng Yao, Jiali Liu, Dakuo Wang, Xiaojuan Ma, Yuxuan Lu, Ying Xu, and Liang He. 2024. Exploring parent's needs for children-centered ai to support preschoolers' interactive storytelling and reading activities. *Proc. ACM Hum.-Comput. Interact.*, 8(CSCW2).

David Rood Traum. 1995. *A computational theory of grounding in natural language conversation*. University of Rochester.

Saranya Venkatraman, Nafis Irtiza Tripto, and Dongwon Lee. 2025. Collabstory: Multi-llm collaborative story generation and authorship analysis. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3665–3679.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, and 1 others. 2024. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345.

Ziyi Wang, Yuxuan Lu, Wenbo Li, Amirali Amini, Bo Sun, Yakov Bart, Weimin Lyu, Jiri Gesi, Tian

Wang, Jing Huang, and 1 others. 2025a. Opera: A dataset of observation, persona, rationale, and action for evaluating llms on human online shopping behavior simulation. *arXiv preprint arXiv:2506.05606*.

Ziyi Wang, Yuxuan Lu, Yimeng Zhang, Pei Chen, Ziwei Dong, Jing Huang, Jiri Gesi, Xianfeng Tang, Chen Luo, and Qun Liu. 2026. Trajectory2task: Training robust tool-calling agents with synthesized yet verifiable data for complex user intents. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 44021–44044.

Ziyi Wang, Yuxuan Lu, Yimeng Zhang, Jing Huang, and Dakuo Wang. 2025b. Customer-R1: Personalized Simulation of Human Behaviors via RL-based LLM Agent in Online Shopping. *Preprint*, arXiv:2510.07230.

Shirley Wu, Michel Galley, Baolin Peng, Hao Cheng, Gavin Li, Yao Dou, Weixin Cai, James Zou, Jure Leskovec, and Jianfeng Gao. 2025. Collabllm: From passive responders to active collaborators. *arXiv preprint arXiv:2502.00640*.

Yu Xia, Jingru Fan, Weize Chen, Siyu Yan, Xin Cong, Zhong Zhang, Yaxi Lu, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2025. AgentRM: Enhancing agent generalization with reward modeling. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 19277–19290, Vienna, Austria. Association for Computational Linguistics.

Ruixuan Xiao, Wentao Ma, Ke Wang, Yuchuan Wu, Junbo Zhao, Haobo Wang, Fei Huang, and Yongbin Li. 2024. Flowbench: Revisiting and benchmarking workflow-guided planning for llm-based agents. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10883–10900.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. 2025. A-mem: Agentic memory for llm agents. *arXiv preprint arXiv:2502.12110*.

Ziqi Yang, Yuxuan Lu, Jennifer Bagdasarian, Vedant Das Swain, Ritu Agarwal, Collin Campbell, Waddah Al-Refaie, Jehan El-Bayoumi, Guodong Gao, Dakuo Wang, Bingsheng Yao, and Nawar Shara. 2026. RECOVER: Designing a Large Language Model-based Remote Patient Monitoring System for Postoperative Gastrointestinal Cancer Care CSCW032. *Proceedings of the ACM on Human-Computer Interaction*, 10(2):1–55.

Bingsheng Yao, Jiaju Chen, Chaoran Chen, April Wang, Toby Jia-jun Li, and Dakuo Wang. 2025a. Through the lens of human-human collaboration: A configurable research platform for exploring human-agent collaboration. *arXiv preprint arXiv:2509.18008*.

Bingsheng Yao, Bo Sun, Yuanzhe Dong, Yuxuan Lu, and Dakuo Wang. 2025b. DPRF: A Generalizable Dynamic Persona Refinement Framework for Optimizing Behavior Alignment Between Personalized LLM Role-Playing Agents and Humans. *Preprint*, arXiv:2510.14205.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. 2024. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. *Preprint*, arXiv:2406.12045.

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

Yimeng Zhang, Tian Wang, Jiri Gesi, Ziyi Wang, Yuxuan Lu, Jiacheng Lin, Simon Zhan, Vianne Gao, Ruochen Jiao, and Junze Liu. 2026. Shop-r1: Rewarding llms to simulate human behavior in online shopping via reinforcement learning. In *International Conference on Learning Representations*, volume 2026, pages 8406–8423.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. 2023. Webarena: A realistic web environment for building autonomous agents. *ArXiv*, abs/2307.13854.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and 1 others. 2024. Sotopia: Interactive evaluation for social intelligence in language agents. In *International Conference on Learning Representations*, volume 2024, pages 40975–41019.

Kunlun Zhu, Hongyi Du, Zhaochen Hong, Xiaocheng Yang, Shuyi Guo, Zhe Wang, Zhenhailong Wang, Cheng Qian, Robert Tang, Heng Ji, and Jiaxuan You. 2025. MultiAgentBench : Evaluating the collaboration and competition of LLM agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8580–8622, Vienna, Austria. Association for Computational Linguistics.

A Mental Model Annotation

A.1 Annotation Schema

Table 5 presents the annotation schema we used to collect ALMANAC, along with the collaboration theories that inform its design.

A.2 Annotation Validation

To examine whether the retrospective annotations retained participant-specific information from the collaboration's corresponding moments, we compared their in-session checkpoint responses with their post-session annotated mental-model rationales. For each checkpoint, we computed its mean SBERT score across all post-session-annotated rationales from the same participant within the corresponding interval (SELF). As a comparison, we computed its similarity with rationales from other participants at the corresponding task position (OTHERS). Table 6 summarizes the results.

Across all conditions, in-session checkpoint responses were more similar to participants' retrospective annotations than to those of other participants. The mean similarity difference was approximately 0.10 overall and remained consistent across the two experimental conditions. These differences were statistically significant ($p = 0.0005$), suggesting that the retrospective annotations preserve participant-specific information reflected in participants' mental models during the collaboration.

B Questionnaire

B.1 Demographic Information Questionnaire Items and Results

Table 7 lists the demographic question items used in our study. Our study included 66 participants. Participants were relatively balanced in gender (32 female, 31 male, 2 non-binary/third gender, and 1 preferred not to disclose). Most participants were between 18–34 years old (28 aged 18–24; 33 aged 25–34), with 4 aged 35–44 and 1 aged 45–54. 34 participants held a master's degree, 24 a bachelor's degree, 4 a high school diploma or equivalent, 2 a doctoral degree, 1 an associate degree, and 1 had less than a high school education.

B.2 Collaboration Style Questionnaire Items

We use items from TeamQ (Britton et al., 2017) to collect participants' perceived collaboration styles. They responded to each item using a 5-point frequency scale. Table 8 shows the items used in our pre-study survey.

C Map Material for Data Collection

We adapted the map materials from Anderson et al. (1991). In the pilot studies, we used the original maps from their work. However, because the original Map Task was conducted through in-person interaction, transferring the task to a computer-mediated setting increased task difficulty and resulted in longer completion times. To better suit our study context, we retained the original map style but reduced the number of landmarks and simplified the route structure. Figures 5 and 6 present the map materials used in our data collection. To standardize the map images for LLM comprehension, we convert each map into a grid-based representation and manually annotate the positions of all landmarks (Figure 7). We then encode the map content in a structured JSON format, which provides the LLM with explicit spatial information about the grid, start point, and landmark regions:

```
{
  "grid_size": [...],
  "start_cell": [...],
  "landmarks": {
    "pyramid": {
      "summary": {
        "bbox": {
          "row_min": ...,
          "row_max": ...,
          "col_min": ...,
          "col_max": ...
        },
        "corners": {
          "top_left": [...],
          "top_right": [...],
          "bottom_left": [...],
          "bottom_right": [...]
        },
        "centroid": [...],
        "boundary_cells": [...]
      },
      "cells": [...],
      "type": "blocked"
    },
    "suspension bridge": {
      ...
    },
    ...
  }
}
```

ID	Category	Question	Response Options		Theoretical Grounding
Q1	Team Goal	At this moment, my partner and I were:	t1	Still figuring out what we needed to do	*Shared Mental Models* (Cannon-Bowers et al., 1993); team task awareness in human–AI teaming (Andrews et al., 2023).
			t2	Working toward a shared understanding	
			t3	Clear on what to do and working on it	
			t4	Something was unclear and we were working it out	
			–	Other	
Q2	Partner Intent	At this moment, I thought my partner:	p1	Understood the situation; on the same page	*Shared Mental Models* (Cannon-Bowers et al., 1993); *Partner Models* and theory of mind in dialogue (Clark and Brennan, 1991; Doyle et al., 2019).
			p2	Probably understood, but I was not fully sure	
			p3	Is waiting for more information	
			p4	Misunderstood; not aligned	
			p5	Gave no clear signal either way	
			–	Other	
Q3	self-reasoning	At this moment, my action was driven by:	r1	Executing an agreed-upon plan	*Grounding acts* and conversational grounding (Clark and Brennan, 1991; Traum, 1995); self-component of *Shared Mental Models* and metacognitive action selection in joint activity (Klein et al., 2005).
			r2	Exploring on my own to gather information	
			r3	Confirming the situation with my partner	
			r4	Grounding – sharing/requesting info to align	
			r5	Repairing a mistake or misunderstanding	
			r6	Waiting for more information	
			–	Other	
Q4	Alignment	At this moment, my partner and I were on the same page (Yes / No). If No, why:	n1	Different understanding of the task goal	*Common ground* (Clark and Brennan, 1991); breakdowns in common ground for joint activity (Klein et al., 2005); coordination breakdowns (Schmidt and Bannon, 1992).
			n2	Different understanding of the current state	
			n3	One of us was missing key information	
			n4	Communication was unclear or ambiguous	
			n5	Technical or interface issue got in the way	

Table 5: Annotation schema used to collect ALMANAC.

Condition	N	$\text{SBERT}_{\text{Self}}$	$\text{SBERT}_{\text{Others}}$	p
Overall	139	0.43	0.33	0.000^*
$C_{\text{not-visible}}$	69	0.42	0.32	0.000^*
C_{visible}	70	0.43	0.33	0.000^*

Table 6: Alignment between participants' in-session checkpoint responses and post-session mental-model annotations. SELF denotes the mean SBERT cosine similarity with the participant's own annotations, while OTHERS denotes similarity with annotations from other participants at the corresponding task position.

Question	Options
What is your gender?	Male; Female; Non-binary / third gender; Prefer not to say
What is your age group?	18–24; 25–34; 35–44; 45–54; 55+
What is the highest level of education you have earned?	Less than high school; High school or equivalent; Associate degree; Bachelor's degree; Master's degree; Doctoral degree

Table 7: Demographic questions used in the pre-study questionnaire.

D Dataset Partition and Distribution

We split ALMANAC at the session level to avoid data leakage across train and test splits. Because the two condition settings change participants' available evidence, we construct splits separately for $C_{not_visible}$ and $C_visible$. We use an approximately 4:1 train/test split within each condition. For $C_{not_visible}$, we assign 9 sessions to training and 3 sessions to test. For $C_visible$, we assign 10 sessions to training and 3 sessions to test, resulting in 19 training sessions and 6 test sessions overall.

To make the training and test sets comparable, we select test sessions through distribution matching rather than random sampling. For each session, we compute role-specific proportions over action types (*draw*, *erase*, *message*, *reset*, and *undo*) and over mental-model labels along three dimensions: team goal, partner intent, and self reasoning. We concatenate the Guide and Follower proportion vectors into a shared session representation, with unseen labels assigned a proportion of zero. Within each condition, we enumerate all candidate subsets

Construct	Item
Task Contribution	Participate actively and accept a fair share of the group work.
Task Contribution	Work skillfully on assigned tasks and complete them on time.
Feedback	Give timely, constructive feedback to team members in the appropriate format.
Communication	Communicate actively and constructively.
Inclusiveness	Encourage all perspectives to be considered and acknowledge contributions of others.
Integration	Constructively build on contributions of others and integrate own work with work of others.
Coordination	Take on an appropriate role in the group, e.g., leader or note taker.
Coordination	Clarify goals and plan the project.
Coordination	Report to team on progress.
Interpersonal Expression	Ensure consistency between words, tone, facial expression, and body language.
Team Climate	Express positivity and optimism about team members and project.
Conflict Management	Display appropriate assertiveness: neither dominating, submissive, nor passive aggressive.
Conflict Management	Contribute appropriately to healthy debate.
Conflict Management	Respond to and manage direct/indirect conflict constructively and effectively.

Table 8: Collaboration tendency questionnaire items used in the pre-study questionnaire.

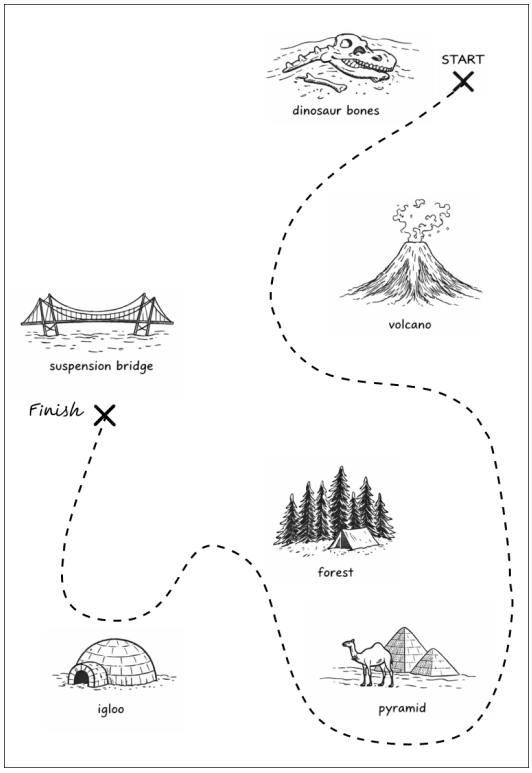

Figure 5: Guide map used in the Map Task. The guide has access to the target route and provides instructions to help the follower reproduce the route.

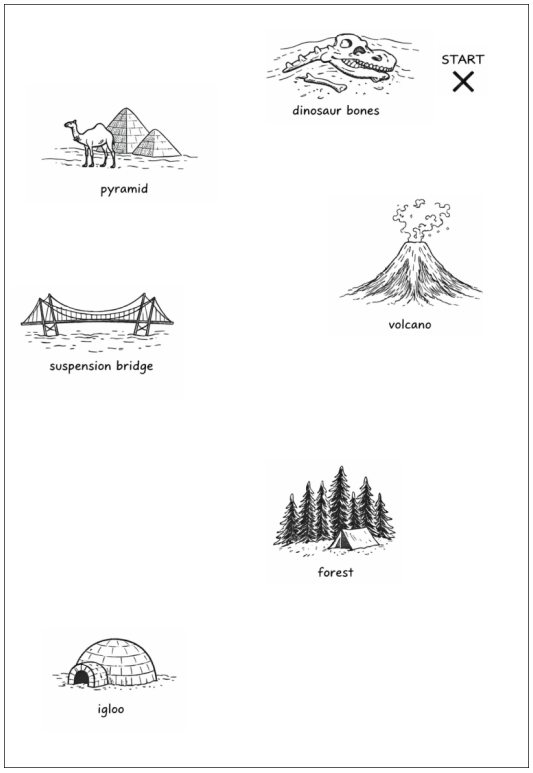

Figure 6: Follower map used in the Map Task. The follower sees the map landmarks but does not have access to the target route.

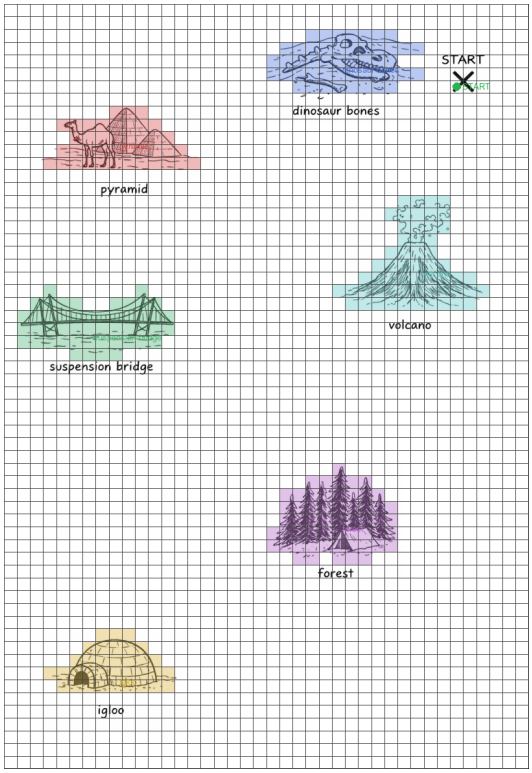

Figure 7: Grid-based map representation used for LLM-readable standardization. Landmark positions are manually annotated and converted into structured spatial representations.

of three test sessions and choose the subset that minimizes the following objective:

$$\mathcal{L} = \sum_{d} \left|\mu_d^{\text{train}} - \mu_d^{\text{test}}\right| + \sum_{d} \left|\sigma_d^{\text{train}} - \sigma_d^{\text{test}}\right|,$$

where d indexes each feature dimension, and μ_d and σ_d denote the across-session mean and standard deviation of session-level proportions. Because the objective compares both central tendency and cross-session variability, the resulting split preserves the overall composition of action types and mental-model labels across training and test sets.

Table 2 reports both the action type and the resulting mental-model distribution. The train and test sets preserve similar label composition across the three dimensions, while retaining lower-frequency labels related to uncertainty, repair, and waiting for information.

E Hyper-Parameters and Experiment Settings

All eight jobs share the same training configuration. We train with a sequence length of 65,536, a global batch size of 32, and a micro-batch size of 1, for 3 epochs in total. We use the Distributed Fused Adam optimizer with adam_beta2 = 0.95, a peak learning rate of 5e-5 (min-lr 0.0) under a cosine annealing schedule with no warmup iterations, and bf16 mixed precision. Checkpoints are saved once per epoch (save-interval = 1), with save-optim disabled. The random seed is fixed to 5678. For training efficiency, we enable sequence parallelism, the distributed optimizer, as well as gradient-reduce and parameter-gather overlap.

We use different parallelism configurations for the two model scales. For Qwen-3-4B, we adopt tensor parallelism (TP) of 4, context parallelism (CP) of 2, and pipeline parallelism (PP) of 1, running on 8 GPUs (1 node). For Qwen3-30B-A3B, we use TP = 4, CP = 2, PP = 4, and expert parallelism (EP) of 2, running on 32 GPUs across 4 nodes.

F Complete Experiment Results

Table 9 reports the complete experimental results for the next action prediction task, additionally including Rouge-L scores and results under Chain-of-Thought prompting. Table 10 reports the complete experimental results for the mental model prediction task, additionally including Rouge-L scores.

G Prompts

Model	$Accuracy_{\text{Action_Type}}$		$Recall_{\text{Action_Type}}$		$SBERT_{\text{Message}}$				$ROUGE_{\text{Message}}$				$Accuracy_{\text{Drawing}}$	
	$C_{visible}$	$C_{not_visible}$	$C_{visible}$	$C_{not_visible}$	$C_{visible}$		$C_{not_visible}$		$C_{visible}$		$C_{not_visible}$		$C_{visible}$	$C_{not_visible}$
	F	F	F	F	G	F	G	F	G	F	G	F	F	F
Qwen3-35B-A3B	0.46	0.54	0.34	0.30	0.22	0.25	0.28	0.27	0.07	0.10	0.09	0.12	0.43	0.45
+*CoT*	0.51	0.53	0.39	0.30	0.21	0.23	0.27	0.27	0.06	0.08	0.07	0.10	0.43	0.38
+*Pred. MM*	0.50	0.53	0.27	0.23	0.21	0.25	0.26	0.26	0.07	0.10	0.08	0.11	0.44	0.45
+*Gold MM*	0.48	0.55	0.38	0.31	0.21	0.27	0.28	0.29	0.06	0.10	0.10	0.13	0.44	0.43
+*CoT + Gold MM*	0.52	0.55	0.39	0.31	0.21	0.25	0.28	0.28	0.06	0.10	0.07	0.11	0.43	0.40
Llama 3.3 70B	0.44	0.51	0.32	0.28	0.22	0.23	0.31	0.27	0.10	0.08	0.12	0.10	0.57	0.45
+*CoT*	0.49	0.44	0.38	0.29	0.21	0.24	0.29	0.27	0.08	0.07	0.10	0.09	0.40	0.43
+*Pred. MM*	0.48	0.51	0.33	0.30	0.26	0.25	0.34	0.27	0.12	0.09	0.14	0.11	0.56	**0.57**
+*Gold MM*	0.51	0.51	0.28	0.30	0.22	0.25	0.30	0.29	0.10	0.10	0.12	0.13	0.46	0.56
+*CoT + Gold MM*	0.51	0.48	0.40	0.31	0.21	0.24	0.29	0.27	0.08	0.08	0.10	0.11	0.35	0.29
GPT-5.5	0.56	0.59	0.44	0.35	0.24	0.32	0.33	0.36	0.09	**0.19**	0.11	0.22	0.55	0.43
+*CoT*	0.58	0.59	**0.48**	0.37	0.25	0.29	0.33	0.33	0.08	0.15	0.10	0.18	0.59	0.42
+*Pred. MM*	0.58	0.60	0.44	0.38	**0.31**	0.33	**0.39**	0.36	**0.14**	0.18	0.17	0.22	0.55	0.48
+*Gold MM*	0.58	**0.61**	0.46	0.37	0.25	0.17	0.33	**0.38**	0.09	0.17	0.11	**0.25**	0.55	0.47
+*CoT + Gold MM*	**0.61**	**0.61**	0.47	**0.40**	0.25	0.34	0.33	0.36	0.08	**0.19**	**0.22**	0.10	**0.60**	0.48
Claude 4.6 Sonnet	0.47	0.54	0.36	0.30	0.23	0.29	0.31	0.31	0.08	0.15	0.10	0.16	0.45	0.44
+*CoT*	0.56	0.57	0.43	0.33	0.23	0.28	0.30	0.31	0.07	0.13	0.08	0.16	0.45	0.43
+*Pred. MM*	0.60	0.53	0.39	0.35	0.29	0.29	0.37	0.32	0.13	0.13	0.16	0.15	0.42	0.38
+*Gold MM*	0.51	0.55	0.39	0.31	0.23	0.31	0.31	0.33	0.08	0.17	0.10	0.18	0.43	0.53
+*CoT + Gold MM*	0.56	0.59	0.41	0.34	0.23	0.32	0.31	0.34	0.07	0.15	0.08	0.18	0.41	0.49
Qwen3-4B Fine-tuned	0.56	0.54	0.37	0.31	0.21	0.35	0.23	0.23	0.07	0.11	0.10	0.06	0.47	0.44
Qwen3-30B-A3B Fine-tuned	0.52	0.52	0.30	0.27	0.20	**0.37**	0.22	0.26	0.06	0.09	0.06	0.06	0.54	0.55

Table 9: Follower Action Type Accuracy and Recall, Message SBERT and ROUGE for Guide (G) and Follower (F) roles, and Follower Drawing Accuracy across six models under $C_{visible}$ and $C_{not_visible}$. *Pred. MM* denotes our two-stage ReAct-inspired baseline, which first predicts the participant's mental model from the observed interaction trajectory and then conditions next-action prediction on the predicted mental model. *Gold MM* denotes conditioning on human-annotated mental models. **Bolded** numbers indicate the best performance for each role and canvas visibility condition.

Model	$Accuracy_{\text{Team_Goal}}$				$Accuracy_{\text{Partner_Intent}}$				$Accuracy_{\text{Self_Reasoning}}$				$SBERT_{\text{Rationale}}$				$ROUGE_{\text{Rationale}}$			
	$C_{visible}$		$C_{not_visible}$		$C_{visible}$		$C_{not_visible}$		$C_{visible}$		$C_{not_visible}$		$C_{visible}$		$C_{not_visible}$		$C_{visible}$		$C_{not_visible}$	
	G	F	G	F	G	F	G	F	G	F	G	F	G	F	G	F	G	F	G	F
Qwen3-35B-A3B	**0.48**	0.69	0.41	0.64	**0.58**	0.73	0.38	0.67	**0.31**	0.56	0.32	0.52	**0.41**	0.46	0.44	0.52	0.14	0.20	0.15	0.20
Llama 3.3 70B	0.43	0.71	**0.56**	0.64	0.38	0.76	0.34	0.72	0.23	0.55	0.31	0.53	**0.41**	0.47	0.44	0.54	0.17	0.21	**0.18**	0.26
GPT-5.5	0.41	0.72	0.35	0.68	0.37	0.75	0.36	0.70	0.29	0.60	0.32	0.59	0.40	0.51	**0.45**	0.55	0.16	0.24	0.17	0.24
Claude 4.6 Sonnet	**0.48**	0.75	0.45	0.68	0.51	0.76	0.45	0.71	0.27	0.56	0.30	0.55	**0.41**	0.52	**0.45**	0.55	0.15	0.25	0.16	0.21
Qwen3-4B Fine-tuned	0.37	**0.81**	0.51	**0.88**	0.40	**0.84**	**0.47**	**0.84**	0.28	**0.65**	0.30	**0.70**	0.37	**0.76**	0.33	0.64	**0.18**	**0.68**	0.16	0.46
Qwen3-30B-A3B Fine-tuned	0.47	0.55	0.39	0.55	0.38	0.78	0.46	0.77	0.29	0.54	**0.33**	0.54	0.34	0.61	0.38	**0.66**	0.17	0.44	**0.18**	**0.55**

Table 10: Team Goal Accuracy, Partner Intent Accuracy, Self-Reasoning Accuracy, Rationale SBERT, and Rationale ROUGE across six models in Guide (G) and Follower (F) roles under $C_{visible}$ and $C_{not_visible}$. **Bolded** numbers indicate the best performance for each role and canvas visibility condition.

G.1 Next Action Prediction

Follower's Prompt

before skip**<Task Description>**
You are participating in a two-player collaborative map-reproduction task. There are two roles: a guide and a follower. The guide can see a map with all landmarks and the correct route. The follower can see a similar map with all landmarks but without the route. The two players need to communicate and coordinate so that the follower can reproduce the guide's route on the follower's map. The task unfolds through a sequence of actions. At each step, the follower may send a message, draw part of the route, erase part of the route, undo the latest edit, or reset the drawing.

<Role Description>
You are role-playing the follower in this task. Your goal is not to solve the task perfectly, but to authentically simulate what this specific human follower would most likely do next, given their persona and the previous interaction history.

<Game Rules>
- You and the guide cannot directly see each other's maps.
- One landmark on your map is misplaced compared with the guide's map. However, you should follow the guide's instructions and reproduce the route on your map.
- You may need to ask clarification questions, acknowledge instructions, draw based on your current understanding, correct previous drawing errors, or wait for more guidance.

<Participant context>
Participant role: `PARTICIPANT_ROLE`
Participant name: `PARTICIPANT_NAME`

`{COLLABORATION_PROFILE}`

Use this profile as a soft behavioral tendency, not a fixed rule. The next mental model should still be primarily grounded in the interaction history and current map state. If the history shows a different behavioral pattern, prioritize the observed interaction history over the TeamQ profile.

<Action Space>
You must choose exactly one next action from the following action types:
message: Send a message to communicate with the guide.
draw: Draw a route segment on the follower's map. The content must be an ordered list of [row, col] cells in the direction of travel.
erase: Erase part of the current drawing. The content must be an ordered list of [row, col] cells to erase.
undo: Undo latest route edit.
reset: Clear entire drawing.

<Map Interpretation>
Discrete grid, 0-based [row, col]. Origin top-left [0, 0]; row increases downward, col increases rightward. Any landmark with "type": "blocked" has a "cells" list, which means those cells are impassable. Your route must NEVER include them. Use bbox / centroid from the landmark reference plus the map image to locate named landmarks. "Bottom / top / left / right" of a landmark refers to that region of the landmark, not the whole map. Paths to a landmark corner usually require BOTH row and col to change — not a single long horizontal or vertical segment.

<Current Map>
`{CURRENT_MAP}`

<Response Format>
{
"action_type": "$message|draw|erase|undo|reset$",
"action_content": "...",
"rationale": "..."
}

For action_content:
- If action_type is "message", action_content must be the exact message text the follower would send.
- If action_type is "draw" or "erase", action_content must be an ordered list of [row, col] cells, for example: [[12, 8], [12, 9], [13, 9]].

- If action_type is "undo" or "reset", action_content must be an empty string "".

For rationale:
Briefly explain why this action is the most likely next action for this follower, based on the persona and interaction history. The rationale should be concise and should not introduce information that is not visible in the input.

<Instructions for aligning with human behaviors>
- Given the interaction history, predict the single next action that this follower would most likely take. Your prediction should be grounded in:
  - the guide's most recent messages;
  - the follower's previous actions and communication style;
  - the follower's persona.
- Do not predict an ideal or optimal action unless it is also likely for this specific follower. Human participants may be incomplete, cautious, redundant, informal, uncertain, or locally focused. Preserve these behavioral patterns when they appear in the history.
- Do not add unnecessary politeness, formal language, or overly detailed explanations unless this follower has shown that style.
- Do not invent information that is not supported by the interaction history.
- Do not mention that you are an AI, a simulator, or making a prediction.

<Interaction History>
`{INTERACTION_HISTORY}`

Guide's Prompt

before skip**<Task Description>**
You are participating in a two-player collaborative map-reproduction task. There are two roles: a guide and a follower. The guide can see a map with all landmarks and the correct route. The follower can see a similar map with all landmarks but without the route. The two players need to communicate and coordinate so that the follower can reproduce the guide's route on the follower's map. The task unfolds through a sequence of actions. At each step, the guide can only send a message.

<Role Description>
You are role-playing the guide in this task. Your goal is not to solve the task perfectly, but to authentically simulate what this specific human guide would most likely do next, given their persona and the previous interaction history.

<Game Rules>
- You and the follower cannot directly see each other's maps.
- One landmark on the follower's map is misplaced compared with your map. You should give instructions so the follower can reproduce the route on their map.
- You may need to ask clarification questions, acknowledge the follower's messages, give route directions based on your understanding of their progress, correct or refine earlier instructions, or wait for more information from the follower.

<Participant context>
Participant role: `PARTICIPANT_ROLE`

Participant name: `PARTICIPANT_NAME`

`{COLLABORATION_PROFILE}`

Use this profile as a soft behavioral tendency, not a fixed rule. The next mental model should still be primarily grounded in the interaction history and current map state. If the history shows a different behavioral pattern, prioritize the observed interaction history over the TeamQ profile.

<Action Space>
The only action type is:
message: Send a message to communicate with the follower.

<Map Interpretation>
Discrete grid, 0-based [row, col]. Origin top-left [0, 0]; row increases downward, col increases rightward. Any landmark with "type": "blocked" has a "cells" list, which means those cells are impassable. Your route must NEVER include them. Use bbox / centroid from the landmark reference plus the map image to locate named landmarks. "Bottom / top / left / right" of a landmark refers to that region of the landmark, not the whole map. Paths to a landmark corner usually require BOTH row and col to change — not a single long horizontal or vertical segment.

<Current Map>
`{CURRENT_MAP}`

<Response Format>
{
  "action_type": "*message*",
  "action_content": "...",
  "rationale": "..."
}

For action_content:
- If action_type is "message", action_content must be the exact message text the guide would send.

For rationale:
Briefly explain why this action is the most likely next action for this guide, based on the persona and interaction history. The rationale should be concise and should not introduce information that is not visible in the input.

<Instructions for aligning with human behaviors>
- Given the interaction history, predict the single next action that this guide would most likely take. Your prediction should be grounded in:
  - the follower's most recent actions;
  - the guide's previous actions and communication style;
  - the guide's persona.
- Do not predict an ideal or optimal action unless it is also likely for this specific guide. Human participants may be incomplete, cautious, redundant, informal, uncertain, or locally focused. Preserve these behavioral patterns when they appear in the history.
- Do not add unnecessary politeness, formal language, or overly detailed explanations unless this follower has shown that style.
- Do not invent information that is not supported by the interaction history.

- Do not mention that you are an AI, a simulator, or making a prediction.

<Interaction History>
{INTERACTION_HISTORY}

G.2 Mental Model Prediction

Follower's Prompt

before skip**<Task Description>**
You are simulating a human participant's action-level mental model in a two-player collaborative map-reproduction task. In this task, there are two roles: guide and follower. The guide can see a map with all landmarks and the correct route. The follower can see a similar map with all landmarks but does not see the route. The two players need to communicate and coordinate so that the follower can reproduce the guide's route on the follower's map. The task unfolds as a sequence of actions. At each step, the follower may send a message, draw part of the route, erase part of the route, undo the latest route edit, or reset the drawing. Your task is to predict the follower's mental model at the current action moment.

<Role Description>
You are simulating the follower's mental model during the map task. Given the participant's persona, the interaction history, the current action, the current drawing state, and any previous mental model annotations, predict what this specific follower would most likely report about:

1. what the team was trying to do;
2. what they thought the guide was trying to do;
3. what they themselves were trying to do;
4. why they took or understood the current action in that way.

The prediction should reflect the follower's subjective understanding at that moment, not the objective ground truth of the task.

<Game Rules>
- The follower and the guide cannot directly see each other's maps.
- One landmark on the follower's map is misplaced compared with the guide's map. However, the follower should follow the guide's instructions and reproduce the route on the follower's map.

<Participant context>
Participant role: PARTICIPANT_ROLE
Participant name: PARTICIPANT_NAME

{COLLABORATION_PROFILE}

Use this profile as a soft behavioral tendency, not a fixed rule. The next mental model should still be primarily grounded in the interaction history and current map state. If the history shows a different behavioral pattern, prioritize the observed interaction history over the TeamQ profile.

<Mental Model Annotation Task>
The participant provided mental model annotations after completing the Map Task. For each action, they were asked to recall what they were thinking at that specific moment. You need to predict the participant's annotation for the current action. The annotation contains four fields:

1. team_goal
What the follower thought the team was trying to do at that moment. Choose exactly one label:
- "Still figuring out what we needed to do" - "Working toward a shared understanding" - "Clear on what to do and working on it" - "Something was unclear and we were working it out" - "Other"

2. partner_intent
What the follower thought the guide was trying to do or understood at that moment. Choose exactly one label:
- "Understood the situation and we were on the same page" - "Probably understood our situation but I was not fully sure" - "Is waiting for more information to understand the situation" - "Misunderstood and we were not aligned" - "Gave no clear signal either way" - "Other"

3. self_reasoning
What the follower thought they themselves were trying to do at that moment. Choose exactly one label:
- "Executing a plan we already agreed on" - "Exploring on my own to gather information" - "Confirming the situation with my partner" - "Grounding by sharing or requesting information to align" - "Repairing a mistake or misunderstanding" - "Waiting for more information" - "Other"

If you select 'Other' for any label, you must provide a specific, meaningful label to replace 'Other'—do not just leave it as 'Other'.

4. rationale
A short free-form explanation, written from the follower's perspective, describing what the team, the guide, and the follower were trying to do at that action moment. The rationale should sound like the participant's own retrospective explanation, not an external analysis.

<Map Interpretation>
Discrete grid, 0-based [row, col]. Origin top-left [0, 0]; row increases downward, col increases rightward. Any landmark with "type": "blocked" has a "cells" list, which means those cells are impassable. Your route must NEVER include them. Use bbox / centroid from the landmark reference plus the map image to locate named landmarks. "Bottom / top / left / right" of a landmark refers to that region of the landmark, not the whole map. Paths to a landmark corner usually require BOTH row and col to change — not a single long horizontal or vertical segment.

<Current Map>
{CURRENT_MAP}

<Current follower action (you are simulating the mental model behind this action)>
{CURRENT_ACTION}

<Response Format>
Return only a valid JSON object. Do not include markdown, explanations, or extra text outside the JSON. The JSON must have exactly the following fields:

{
"team_goal": "...",
"partner_intent": "...",
"self_reasoning": "...",
"rationale": "..."
}
The three label fields must exactly match one of the allowed labels.

<Instructions for aligning with human behaviors>
- Given the interaction history and previous mental models, predict this follower's mental model for the current action. Your prediction should be grounded in:
  - the guide's most recent messages;
  - the follower's previous actions and communication style;
  - the follower's previous mental models;
  - the follower's persona.
- Follow the participant's previous behaviors/habits (e.g., writing styles, preferences, etc.) when reporting their mental model.
- Do not add unnecessary politeness, formal language, or overly detailed explanations unless this follower has shown that style.
- Write the rationale in a first-person perspective, as if the participant is recalling their own thought process after the task.
- Do not invent information that is not supported by the interaction history.
- Do not mention that you are an AI, a simulator, or making a prediction.

<Interaction History>
`{INTERACTION_HISTORY}`

Guide's Prompt

before skip**<Task Description>**
You are simulating a human participant's action-level mental model in a two-player collaborative map-reproduction task. In this task, there are two roles: guide and follower. The guide can see a map with all landmarks and the correct route. The follower can see a similar map with all landmarks but does not see the route. The two players need to communicate and coordinate so that the follower can reproduce the guide's route on the follower's map. The task unfolds as a sequence of actions. At each step, the guide will send a message. Your task is to predict the guide's mental model at the current action moment.

<Role Description>
You are simulating the guide's mental model during the map task. Given the participant's persona, the interaction history, the current action, and any previous mental model annotations, predict what this specific guide would most likely report about:

1. what the team was trying to do;
2. what they thought the guide was trying to do;
3. what they themselves were trying to do;
4. why they took or understood the current action in that way.

The prediction should reflect the guide's subjective understanding at that moment, not the objective ground truth of the task.

<Game Rules>
- The follower and the guide cannot directly see each other's maps.
- One landmark on the follower's map is misplaced compared with the guide's map. However, the follower should follow the guide's instructions and reproduce the route on the follower's map.

<Participant context>
Participant role: `PARTICIPANT_ROLE`
Participant name: `PARTICIPANT_NAME`

`{COLLABORATION_PROFILE}`

Use this profile as a soft behavioral tendency, not a fixed rule. The next mental model should still be primarily grounded in the interaction history and current map state. If the history shows a different behavioral pattern, prioritize the observed interaction history over the TeamQ profile.

<Mental Model Annotation Task>
The participant provided mental model annotations after completing the Map Task. For each action, they were asked to recall what they were thinking at that specific moment. You need to predict the participant's annotation for the current action. The annotation contains four fields:

1. team_goal
What the guide thought the team was trying to do at that moment. Choose exactly one label:
- "Still figuring out what we needed to do" - "Working toward a shared understanding" - "Clear on what to do and working on it" - "Something was unclear and we were working it out" - "Other"

2. partner_intent
What the guide thought the guide was trying to do or understood at that moment. Choose exactly one label:
- "Understood the situation and we were on the same page" - "Probably understood our situation but I was not fully sure" - "Is waiting for more information to understand the situation" - "Misunderstood and we were not aligned" - "Gave no clear signal either way" - "Other"

3. self_reasoning
What the guide thought they themselves were trying to do at that moment. Choose exactly one label:
- "Executing a plan we already agreed on" - "Exploring on my own to gather information" - "Confirming the situation with my partner" - "Grounding by sharing or requesting information to align" - "Repairing a mistake or misunderstanding" - "Waiting for more information" - "Other"

If you select 'Other' for any label, you must provide a specific, meaningful label to replace 'Other'—do not just leave it as 'Other'.

4. rationale
A short free-form explanation, written from the follower's perspective, describing what the team, the guide, and the follower were trying to do at that action moment. The rationale should sound like the participant's own retrospective explanation, not an external analysis.

<Map Interpretation>
Discrete grid, 0-based [row, col]. Origin top-left [0, 0]; row increases downward, col increases rightward. Any landmark with "type": "blocked" has a "cells" list, which means those cells are impassable. Your route must NEVER include them. Use bbox / centroid from the landmark reference plus the map image to locate named landmarks. "Bottom / top / left / right" of a landmark refers to that region of the landmark, not the whole map. Paths to a landmark corner usually require BOTH row and col to change — not a single long horizontal or vertical segment.

<Current Map>
{CURRENT_MAP}

<Current guide action (you are simulating the mental model behind this action)>
{CURRENT_ACTION}

<Response Format>
Return only a valid JSON object. Do not include markdown, explanations, or extra text outside the JSON. The JSON must have exactly the following fields:

```
{
    "team_goal": "...",
    "partner_intent": "...",
    "self_reasoning": "...",
    "rationale": "..."
}
```

The three label fields must exactly match one of the allowed labels.

<Instructions for aligning with human behaviors>
- Given the interaction history and previous mental models, predict this guide's mental model for the current action. Your prediction should be grounded in:
    - the follower's most recent actions;
    - the guide's previous actions and communication style;
    - the guide's previous mental models;
    - the guide's persona.
- Follow the participant's previous behaviors/habits (e.g., writing styles, preferences, etc.) when reporting their mental model.
- Do not add unnecessary politeness, formal language, or overly detailed explanations unless this follower has shown that style.
- Write the rationale in a first-person perspective, as if the participant is recalling their own thought process after the task.
- Do not invent information that is not supported by the interaction history.
- Do not mention that you are an AI, a simulator, or making a prediction.

<Interaction History>
{INTERACTION_HISTORY}