Embodied Multi-Agent Coordination by Aligning World Models Through Dialogue

Vardhan Dongre & Dilek Hakkani-Tür
Siebel School of Computing & Data Science
University of Illinois Urbana-Champaign
{vdongre2, dilek}@illinois.edu

Abstract

Effective collaboration between embodied agents requires more than acting in a shared environment; it demands communication grounded in each agent's evolving understanding of the world. When agents can only partially observe their surroundings, coordination without communication is provably hard, but communication can, in principle, bridge this gap by allowing agents to share observations and align their world models. In this work, we examine whether LLM-based embodied agents actually realize the ability to communicate. We extend PARTNR, a benchmark for collaborative household robotics, with a natural-language dialogue channel that enables two agents with partial observability to communicate during task execution. To evaluate whether dialogue leads to genuine world-model alignment rather than superficial coordination, we propose a framework for measuring world-model alignment defined over per-agent world graphs: observation convergence (do private world models align over time?), information novelty (do messages convey what the partner lacks?), and belief-sensitive messaging (do agents model what their partner knows?). Our experiments across three LLMs reveal that dialogue reduces action conflict by 41–93 percentage points but degrades task success relative to silent coordination. Using our metrics, we characterize the gap between superficial coordination and genuine world-model alignment, and identify where current models fall on this spectrum.

1 Introduction

When humans coordinate on a shared physical task, they naturally benefit from language. They negotiate responsibilities, share what they observe, flag unexpected obstacles, and revise their joint plan as new information arrives (Clark, 1996; Tomasello et al., 2005). Every utterance carries an implicit model of the listener: what they already know, what they need to hear, and what they will do next. Effective collaboration is not about exchanging information; it is about using language to **align world models** so that two decentralized minds can plan as though they shared one. Central to this process is Theory of Mind (Premack and Woodruff, 1978) where each partner continuously models what the other knows, doesn't know, what they intend, and uses language to close the gaps (Grosz and Kraus, 1996; Frank and Goodman, 2012). Therefore, in problems grounded in a physical (or simulated) world, effective joint planning is not about exchanging arbitrary information; it is about aligning individual world models so that decentralized agents can plan as one.

This intuition has a precise formal counterpart in joint planning with multiple decentralized agents. Decentralized partially observable Markov decision processes (Dec-POMDPs) with two or more agents are provably intractable (NEXP-complete) (Bernstein et al., 2002), a double-exponential jump over the single-agent case. But with an effective communication channel, the problem collapses towards centralized planning, where a single coordinator sees all observations (which is only PSPACE-complete) [1] (Goldman and Zilberstein, 2004, 2008). The reduction happens because a shared communication channel lets two agents' partial observations be combined into a joint view; reducing the problem to its centralized form, where a single planner has access to both agents' observations.

Large language models now routinely serve as planners for embodied agents (Zhang and Soh, 2023; Mandi et al., 2024; Zu et al., 2025; Gong et al., 2024; Dongre et al., 2025), and on text-based benchmarks they reportedly approach human per-

[1] NEXP-complete problems require time exponential in an exponential function of input size in the worst case; PSPACE-complete problems require only polynomial memory. The gap quantifies how much harder two-agent decentralized planning is than single-planner planning; communication is, in principle, how the gap is closed.

Proceedings of the 27th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 293–306
August 2–5, 2026.

formance on Theory of Mind tasks (Kosinski, 2024; Strachan et al., 2024; Street et al., 2024), though this is contested (Shapira et al., 2024). Yet these capabilities have not been evaluated together. Embodied multi-agent setups report task success without probing whether communication is meaningful, with few exceptions (Sclar et al., 2023; Ma et al., 2023; Bara et al., 2021). We study natural-language dialogue for two reasons. First, LLM-based embodied agents communicate through text tokens, the only output channel their decoder produces. Second, humans coordinate on shared physical tasks through natural language under the same partial-observation conditions. We therefore ask: when two embodied LLM agents coordinate through dialogue in a physically grounded environment where miscommunication costs real steps, do they attempt to align their world models, or do they merely appear to?

We study this question in the PARTNR benchmark for collaborative household robotics (Szot et al., 2024), extended with a dialogue channel that allows two LLM agents with partial observability grounded in the household environment to exchange free-form natural-language messages during task execution. Our contributions are as follows:

- A diagnostic framework for measuring world-model alignment in embodied dialogue. We formalize alignment as the convergence between agents' private world graphs and propose diagnostic metrics that operationalize distinct aspects of dialogue effectiveness defined from a formal per-agent world graph.
- An empirical study of embodied LLM dialogue under varied communication architectures and resource constraints, one where messaging competes with motor actions for the same step budget and one where messaging is free, and across prompt policies that vary the agents' default messaging behavior.

2 Task Design and Setup

We build on the PARTNR benchmark (Szot et al., 2024), a platform for evaluating LLM-based planners on collaborative household robotics tasks in the Habitat simulator (Puig et al., 2024). A task pairs a 3D scene drawn from the Habitat Synthetic Scenes Dataset (HSSD) (Khanna et al., 2024) with a free-form natural-language instruction (e.g., *"Place the green apple on the kitchen counter and put the book on the shelf"*) and a set of ground-truth evaluation predicates that determine success.

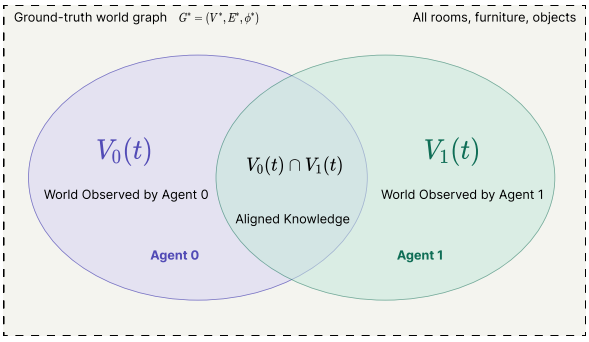

Figure 1: Partial observability in the two-agent setting. Each agent directly observes a private subset $V_i(t)$ of the ground-truth scene $G^\star$; their intersection $V_0 \cap V_1$ is aligned knowledge, and the complement of their union is unobserved by both. Observation convergence OC (§3.3) is the Jaccard overlap of $V_0(t)$ and $V_1(t)$.

Agents and skills: Two heterogeneous embodied agents, a quadruped robot and a humanoid, share the scene. Each is controlled by an independent LLM planner that selects **high-level actions** from a shared API: `Navigate`, `Pick`, `Place`, `Open`, `Close`, `Rearrange`, and object-state actions such as `Clean`, `Fill`, `Pour`, and `PowerOn/Off` (see Table 1). Each high-level action is executed by a **low-level motor skill** (Szot et al., 2024); we use oracle skills that deterministically ground each action in the simulator, so that coordination, not motor control, is the subject of study. The two action spaces are deliberately **asymmetric**: state-change actions (`Clean`, `Fill`, `PowerOn/Off`, `Pour`) are available only to the humanoid. Tasks therefore cannot be solved by either agent alone: the robot must locate and stage objects that the humanoid then modifies, creating coordination demands that, in principle, dialogue is well-suited to address. Every action (motor, perception, or, in our setting, communication) consumes one step from a shared budget T, which in our experiments is fixed at $T = 10{,}000$ simulator steps per episode.

Private world graphs: Each agent maintains a private world graph: a hierarchical representation of the scene as a tree $\{house \rightarrow room \rightarrow furniture \rightarrow object\}$. The graph is populated incrementally as the agent explores; rooms the agent has not yet visited do not appear, and objects inside unopened receptacles are hidden until observed. The two agents are spawned in different rooms with

Category	**Agent 0:** Robot	**Agent 1:** Humanoid
Navigation	`Navigate`, `Explore`	`Navigate`, `Explore`
Manipulation	`Pick`, `Place`, `Rearrange`, `Open`, `Close`	`Pick`, `Place`, `Rearrange`, `Open`, `Close`
State-change	*not available*	`Clean`, `Fill`, `Pour`, `PowerOn/Off`,
Perception	`FindObject`, `FindReceptacle`, `FindRoom`, `FindAgent`	`FindObject`, `FindReceptacle`, `FindRoom`, `FindAgent`
Control	`Wait`, `Done`	`Wait`, `Done`
SC (synchronous, costed):		
Dialogue	`SendMessage`, `ReadMessages`	`SendMessage`, `ReadMessages`
ACF (asynchronous, free):		
Dialogue	`SendMessage` (dual-output)	`SendMessage` (dual-output)

Table 1: Per-agent action space. The two agents share most capabilities, but *state-change* actions (highlighted) are exclusive to the humanoid (Agent 1). The Silent baseline omits both dialogue rows. SC adds explicit `SendMessage` and `ReadMessages` tools, each of which costs one planner step. ACF replaces these with a dual-output response format (thought, message, action) in which messaging does not consume a step and partner messages auto-inject into observations.

asymmetric partial observability: their initial world graphs are disjoint, and neither has access to the other's. This introduces the core coordination challenge: for any object outside the agent's own known region, the agent must either explore (expensive) or rely on the partner to have seen it (requires communication).

Dialogue channel: To study world-model alignment, we introduce a shared **dialogue channel** which is a message buffer visible to both agents' planners. The channel exposes two tools that sit alongside the motor skills in each agent's tool schema:

- `SendMessage(content)`, append a free-form natural-language message to the buffer, attributed to the sender and timestamped with the current simulator step. Executing this tool consumes one step from the action budget.
- `ReadMessages()`, return all messages in the buffer since the caller's last read. This tool does not consume a step, consistent with other perception tools.

Partner messages are injected into the planner context at the next replan, interleaved with the agent's own action history in ReAct format (Yao et al., 2023). The agents otherwise share nothing: world graphs remain private, and the only information flowing between them is through dialogue and shared observable space.

Communication Architectures: The dialogue channel admits two distinct communication architectures that determine how speaking interacts with acting, and we evaluate both. In the Synchronous-Costed (SC) architecture, a natural extension of PARTNR's action space, `SendMessage` is one of the agent's available actions at each replan: choosing to send a message means *not* choosing to act (motor) in the environment that turn. Speaking competes with acting for the same step budget. In the Asynchronous-Cost-Free (ACF) architecture, the agent's planner emits a structured response containing both an optional message and a required motor action in a single LLM call; messages ride along with actions and do not consume the action budget, and partner messages are auto-injected into observations at every replan boundary (removing the need for an explicit `ReadMessages` call). The two architectures isolate a fundamental design question for embodied conversational agents: whether reasoning, speaking, and acting must be serialized, or can be harmonized within a single decision (Dongre et al., 2025). We use the SC vs. ACF contrast in §5 to test whether action-budget contention is the mechanism behind dialogue's effect on task success.

Planning loop: Planners operate in the ReAct and ReSpAct paradigm (Yao et al., 2023; Dongre et al., 2025). At each planning step, the agent's LLM receives a prompt containing the task instruction, the current world graph serialized as text, a tool schema, and the history of its own prior thoughts and actions. The LLM emits a free-form *thought* followed by a structured *action*, which is parsed and dispatched to the corresponding tool or motor skill. Execution returns an observation, which is appended to the history, and the loop repeats until the task is solved or the step budget is exhausted. Of the tool categories in Table 1, only motor and control actions (navigation, manipulation, articulation, state-change) consume steps

from the budget; perception tools (`FindObject`, `FindReceptacle`, `FindRoom`, `FindAgent`) query the agent's world graph without advancing the simulator and do not count against the budget.

3 World Models, Alignment, and Evaluation

Our analysis utilizes a single formal object, the per-agent world graph from which task success, conflict, and the diagnostic metrics of dialogue effectiveness are all derived. We define the object, then the success and conflict measures, then introduce two notions of alignment (observation-based and belief-based) and metrics that operationalize them.

3.1 Formal Setup

The scene at any simulator step is a labeled tree $G^\star = (V^\star, E^\star, \phi^\star)$, where $V^\star$ is the set of entities (rooms, furniture, objects), $E^\star$ is the containment hierarchy *house* $\rightarrow$ *room* $\rightarrow$ *furniture* $\rightarrow$ *object*, and $\phi^\star$ assigns attributes (entity type and state, e.g. `clean`, `filled`, `powered_on`). Each agent $i \in \{0, 1\}$ maintains two entity sets. $V_i(t) \subseteq V^\star$ is the observation set: entities agent i has directly observed by step t. $B_i(t) \supseteq V_i(t)$ is the belief set: entities agent i has either directly observed or been informed about by its partner. Both sets grow monotonically as agents explore and exchange messages, and they coincide when no dialogue has occurred ($B_i(t) \equiv V_i(t)$). Agents are spawned in different rooms with asymmetric partial observability; neither agent has direct access to the other's V or B. The only channel for sharing entity knowledge is dialogue.

To connect messages to these sets, we define the **entity mention set** $\mu(m) \subseteq V^\star$ for a message m: the entities referenced by simulator handle in the message text, extracted via deterministic pattern matching, with an LLM-judge sensitivity check on a stratified sample (Appendix D). Belief sets are extended through the partner's mentions:

$$B_i(t) \;=\; V_i(t) \;\cup \bigcup_{m \in \mathrm{rcv}_i(t)} \mu(m), \qquad (1)$$

where $\mathrm{rcv}_i(t)$ is the set of messages received by agent i by step t.

Figure 1 illustrates $V_0(t)$ and $V_1(t)$ within $G^\star$. Agents' V-sets are generally overlapping (shared observation of common rooms), generally subset-proper to $G^\star$ (neither has seen everything), and asymmetrically initialized (the initial intersection is the static room layout, not movable objects).

3.2 Task Success and Conflict

We adopt the PARTNR evaluation framework (Szot et al., 2024) unchanged. An instruction $\mathcal{L}$ compiles to an evaluation object $\mathcal{E} = \langle P, C \rangle$ with propositions $P = \{p_1, \ldots, p_n\}$ over world state and constraints $C = \{c_1, \ldots, c_m\}$ over their satisfaction history. For example, the instruction "first, put the toy pineapple next to the toy food; then move them to the shelves" compiles to propositions p_1 =`is_next_to(toy_pineapple, toy_food)` and p_2 =`is_on(toy_pineapple, shelves)`, and constraint c_1 =`before(`p_1`,` p_2`)` (temporal ordering).

Let $\mathcal{W}_t$ denote the ground-truth world state (the scene tree $G^\star$ with current attributes $\phi^\star$) at step t, against which each proposition is evaluated, $p_i(\mathcal{W}_t) \in \{0, 1\}$; $\mathcal{W}_{1:T}$ is the episode's state trajectory. Let $\tau_i = \min\{t : p_i(\mathcal{W}_t) = 1\}$ be the first-satisfaction time of p_i ($\tau_i = \infty$ if never satisfied). Proposition p_i is *complete* if it is both satisfied and constraint-valid:

$$\mathbb{1}_i^{\mathrm{done}} \;=\; \mathbb{1}[\tau_i \neq \infty] \;\cdot\; \prod_j c_j(\tau_{1:n}, \mathcal{W}_{1:T})_i. \quad (2)$$

Here $c_j(\cdot)_i \in \{0, 1\}$ is the verdict of constraint c_j restricted to p_i: it is 0 if c_j involves p_i and is violated, and 1 otherwise (including when c_j does not involve p_i). We report two scores: *percent complete* (PC), the fraction of propositions completed in the episode, and *success rate* (SR), the indicator that all propositions were completed:

$$\mathrm{PC} \;=\; \tfrac{1}{n} \sum_{i=1}^{n} \mathbb{1}_i^{\mathrm{done}}, \qquad \mathrm{SR} \;=\; \mathbb{1}[\mathrm{PC} = 1]. \qquad (3)$$

Beyond task completion, we measure *action conflict*: a joint decision that produces redundant or incompatible effort, where coordination would have achieved the same task progress with strictly fewer resources. Let $\mathcal{T}$ be the set of steps at which both agents replan, and let

$$\alpha_t^{(k)} = (\mathtt{tool}_t^{(k)}, \mathtt{arg}_t^{(k)})$$

denote agent k's action at step t, where `arg` is the target entity handle. We define conflict rate as

$$\mathcal{R}_{\mathrm{conf}} \;=\; \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \mathbb{1}\left[\alpha_t^{(0)} = \alpha_t^{(1)}\right] \;\in\; [0, 1].$$

3.3 Diagnostic Metrics

We define metrics organized into two pairs. The first pair (OC, BC) characterizes alignment of the agents' *world models*; the derived quantity Δ_{align} isolates dialogue's contribution beyond co-exploration. The second pair (IN, BSM) characterizes the *dialogue messages themselves*.

Observation Convergence (OC): **OC** measures the extent to which the agents have directly observed the same entities. It is the Jaccard similarity of the agents' observation sets V_0, V_1:

$$\text{OC}(t) \;=\; \frac{|V_0(t) \cap V_1(t)|}{|V_0(t) \cup V_1(t)|}. \tag{4}$$

OC does not credit dialogue: an agent told about an entity it has not seen does not have that entity in V_i. OC may in fact *drop* under dialogue if one agent skips exploring a region the partner has already described.

Belief Convergence (BC) and Alignment Gap: BC is the Jaccard similarity of the agents' belief sets B_0, B_1 (entities each agent has either directly observed or been informed about by its partner):

$$\text{BC}(t) \;=\; \frac{|B_0(t) \cap B_1(t)|}{|B_0(t) \cup B_1(t)|}. \tag{5}$$

The alignment gap measures:

$$\Delta_{\text{align}}(t) \;=\; \text{BC}(t) - \text{OC}(t) \tag{6}$$

which isolates dialogue's contribution to alignment beyond shared exploration. Silent baselines give $\Delta_{\text{align}} = 0$ by construction. We additionally report $\Delta_{\text{align}}^{\text{grounded}}$, computed with hallucinated handles excluded from $\mu(m)$ (restricting mentions to $V_s(t) \cup V_r(t)$); the gap between the two variants quantifies the alignment cost hallucinated content imposes.

Information Novelty (IN): This measures the per-message informativeness of dialogue relative to the receiver's *direct observations*. For a message m with sender s, receiver r, and time t:

$$\text{IN}(m) \;=\; \frac{|\mu(m) \cap (V_s(t) \setminus V_r(t))|}{|\mu(m)|}. \tag{7}$$

IN uses the observation set V_r rather than the belief set B_r in the denominator-complement: a mention is novel if the receiver has not *seen* the referenced entity, regardless of whether the receiver has been told about it before. This measures whether dialogue transfers new observations rather than repeating what the receiver has already heard. Each mentioned handle falls into one of four categories (illustrated in Figure 2): *both* ($h \in V_s \cap V_r$, common ground), *only-sender* ($V_s \setminus V_r$, novel transfer), *only-receiver* ($V_r \setminus V_s$, sender uninformed), or *neither* ($h \notin V_s \cup V_r$, hallucinated or forecasted).

Belief-Sensitive Messaging (BSM): BSM quantifies whether a sender chooses to mention information that is *new to the receiver*. At each time step t, let

$$\Delta_r(t) = V_s(t) \setminus V_r(t)$$

denote the set of entities known to the sender but not to the receiver. Intuitively, BSM compares how often *mentioned* entities are receiver-novel against how often receiver-novel entities appear among all *mentionable* entities (i.e., those known to the sender).

Formally, let $p_{\text{ment}} = \Pr(v \in \Delta_r \mid v \in \mu(m))$ be the probability that a mentioned entity is receiver-novel, and $p_{\text{base}} = \Pr(v \in \Delta_r \mid v \in V_s)$ the corresponding base rate over all entities known to the sender. BSM is their difference:

$$\text{BSM} \;=\; p_{\text{ment}} - p_{\text{base}}. \tag{8}$$

A positive value ($\text{BSM} > 0$) indicates *targeted information transfer*: the sender preferentially mentions entities the receiver does not yet know. A negative value ($\text{BSM} < 0$) indicates a bias toward *common ground*, reflecting grounding or coordination behavior rather than a lack of partner modeling. We stratify BSM by the sender's *speech-act tag*, an intent label the prompt asks each agent to prepend to its message, drawn from {PLAN, STATUS, CONFIRM, BLOCKED, CORRECT}.

4 Experimental setup

We evaluate the framework along two axes. The **model-robustness** axis asks whether the patterns we identify hold across LLMs of different capability classes. The **communication-architecture** axis asks how dialogue's effect on task success varies across qualitatively different ways of structuring the communication channel: no channel (*Silent*), the Synchronous-Costed dialogue setup (*SC*), and the Asynchronous-Cost-Free dialogue setup (*ACF*; see §2).

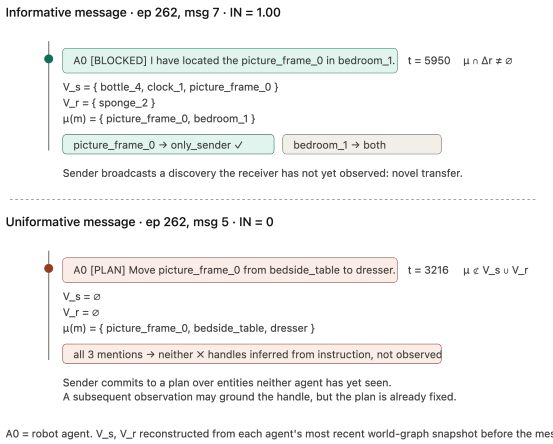

Figure 2: Per-message diagnosis via the four-way handle classification. Each mentioned handle $h \in \mu(m)$ is classified as *only-sender* ($h \in V_s \setminus V_r$), *only-receiver* ($h \in V_r \setminus V_s$), *both* ($h \in V_s \cap V_r$), or *neither* ($h \notin V_s \cup V_r$). **Top:** an informative message whose mention of `picture_frame_0` is novel to the receiver, yielding $\mathrm{IN}(m) = 1.00$. **Bottom:** a message whose handles are neither observed by sender nor receiver at mention time, yielding $\mathrm{IN}(m) = 0$ and instantiating $\mu(m) \not\subseteq V_s \cup V_r$.

Models: We use three LLMs as planners, drawn from the Anthropic and Mistral families: Claude 3.5 Sonnet, Claude 3.5 Haiku, and Mistral Large. Sonnet serves as the headline model for the architectural comparison; the multi-model robustness comparison runs on all three. Decoding parameters are PARTNR defaults throughout.

Conditions: On Sonnet we compare five conditions crossing architecture (Silent, SC, ACF) with prompt policy (*sparingly* vs. *actively encouraged*). Silent agents have no access to the dialogue channel. SC (*sparingly*) is the standard dialogue setup with the fewshot prompt that instructs agents to message only when necessary; SC (*encouraged*) replaces the prompt with one that invites more frequent messaging. ACF (*sparingly*) and ACF (*encouraged*) apply the same two prompts to the Asynchronous-Cost-Free architecture. The five-condition design separates the effect of the architecture from the effect of the prompt policy and supports the mechanism argument in §6. For the multi-model robustness comparison, we run Silent, SC, and ACF on Haiku and Mistral using the *sparingly* prompt only.

Episodes: We use a fixed 100-episode subset of the PARTNR validation set, stratified across the four task families (rearrangement, spatial-constraint, temporal/sequential, object-state) and fixed in advance of any condition being run. For the Sonnet conditions we run all 100 episodes per condition; 20 episodes are dropped from paired comparisons because they failed to reach a termination event in at least one dialogue condition, leaving $N{=}80$ in the intersected slate. Episode exclusion is condition-symmetric: any episode dropped for one condition is dropped from all conditions in the corresponding comparison.

Step budget and constants: All episodes use a $T{=}10{,}000$ simulator-step budget, oracle motor skills (so coordination, not motor control, is the studied variable), partial observability with no agent privileged with ground-truth scene access, and asymmetric agent spawning (see §2). Each agent's planner emits a thought and an action at every replan boundary. Replanning is triggered by skill completion, observation change, or partner-message arrival.

Metrics reported: All reported SR, PC, and $\mathcal{R}_{\text{conf}}$ values are means over episodes of the per-episode quantities defined in §3.2. We report task success rate (SR) and percent-complete (PC) as defined in §3.2; conflict rate $\mathcal{R}_{\text{conf}}$ as a coordination measure; mean simulator-steps and mean dialogue turns as efficiency measures; and the four diagnostic metrics from §3.3: OC, BC and the alignment gap Δ_{align} (together with its hallucination-excluded variant $\Delta_{\text{align}}^{\text{grounded}}$), per-message IN, and BSM both pooled and stratified by speech-act intent. Paired condition-deltas (ΔSR, ΔPC, $\Delta\mathcal{R}_{\text{conf}}$) are computed over the intersection of episode IDs that terminated in all compared conditions.

5 Results

We present results in two layers. §5.1 establishes the conflict–success tradeoff and verifies it across models. §5.2 uses the diagnostic metrics introduced in §3.3 to localize the failure to dialogue content, showing that hallucinated references dominate whenever the channel is used.

5.1 The Conflict-vs-Success Tradeoff

Synchronous-Costed (SC; §2) dialogue substantially reduces action conflict but degrades task success across all evaluated models. On Sonnet ($N{=}100$; Table 2), SC reduces $\mathcal{R}_{\text{conf}}$ (defined in §3.2) by 93 percentage points relative to silent agents, but lowers success rate (SR) by 17 points.

Condition	SR	PC	$\mathcal{R}_{\text{conf}}$	turns/ep
Silent	0.706	0.755	0.941	0.00
SC	0.529	0.716	0.015	3.76
SC*	0.471	0.615	0.000	4.18
ACF	0.412	0.520	0.224	13.35
ACF*	0.392	0.410	0.014	19.15

Table 2: Sonnet on the headline paired slate (N=80 episodes present in all conditions). Silent agents have no dialogue channel; SC is the standard dialogue setup (*sparingly* prompt by default); SC^* uses the *actively encouraged* dialogue prompt under the same Synchronous-Costed architecture; ACF uses an Asynchronous-Cost-Free architecture with a structurally-encouraged policy that signals the channel as a first-class action surface.

Model	ΔSR	ΔPC	$\Delta\mathcal{R}_{\text{conf}}$
Sonnet	-0.177	-0.039	-0.926
Haiku	-0.368	-0.331	-0.602
Mistral-L	-0.050	-0.012	-0.405

Table 3: Paired SC-minus-silent deltas across three LLMs. The conflict-vs-success tradeoff replicates across all three models; for no model does SC dialogue produce a positive ΔSR.

The Asynchronous-Cost-Free (ACF) architecture retains a 72-point reduction in conflict but *does not recover task success*. ACF also underperforms the silent baseline, indicating that removing the action cost of communication alone is insufficient to restore performance.

This pattern is consistent across models on an 80-episode subset (Table 3): ΔSR is negative for every model ($\Delta\text{SR} \in [-0.37, -0.05]$, $p < 0.01$, paired Wilcoxon), and $\Delta\mathcal{R}_{\text{conf}}$ is uniformly negative ($\Delta\mathcal{R}_{\text{conf}} \in [-0.93, -0.41]$). In our evaluated conditions, dialogue does not improve task success.

Removing the cost of messaging increases communication volume by $\sim$3.5$\times$ (13.4 vs. 3.8 turns per episode), but does not yield a practically meaningful improvement in SR (ΔSR = -0.11, $p = 0.01$). Encouraging more dialogue within SC similarly fails to improve SR (ΔSR = -0.05, $p = 0.006$). The degradation in task success is therefore robust across the communication interventions we test.

5.2 World-Model Alignment Under Each Condition

The diagnostic metrics from §3.3 explain the SR degradation. Table 4 reports end-of-episode values; Figure 3 shows trajectories.

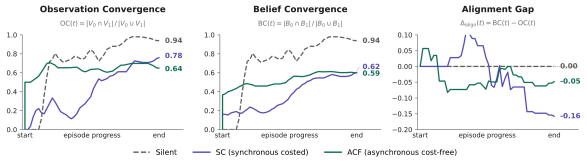

Figure 3: Trajectories of observation convergence (OC), belief convergence (BC), and Δ_{align}.

Condition	OC	BC	Δ_{align}	$\Delta_{\text{align}}^{\text{grounded}}$
Silent	0.92	0.92	0.00	0.00
SC	0.67	0.51	-0.15	$+0.19$
SC*	0.72	0.57	-0.15	$+0.17$
ACF	0.65	0.61	-0.05	$+0.06$
ACF*	0.75	0.63	-0.07	$+0.12$

Table 4: End-of-episode alignment metrics on the Sonnet model. Higher OC and BC indicate greater world-model overlap. For Δ_{align}, positive values indicate dialogue adds alignment, negative values indicate dialogue reduces it (relative to silent co-exploration). The grounded variant excludes hallucinated handles from the mention set; pooled Δ_{align} is negative for every dialogue condition, but $\Delta_{\text{align}}^{\text{grounded}}$ is positive for every dialogue condition, grounded dialogue would align beliefs as the formalism predicts

Reading the metrics: OC and BC are Jaccard overlaps; higher values indicate greater agreement between agents. Under silent, $B_i \equiv V_i$, so BC = OC and $\Delta_{\text{align}} = 0$. Positive Δ_{align} indicates alignment beyond co-exploration; negative values indicate misalignment. The grounded variant excludes hallucinated references.

Hallucinations cancel alignment: Pooled Δ_{align} is negative in all dialogue conditions (-0.15 SC, -0.10 SC*, -0.05 ACF; $p \approx 0.002$), while $\Delta_{\text{align}}^{\text{grounded}}$ is positive ($+0.19$, $+0.17$, $+0.06$). Thus, grounded dialogue aligns beliefs, but hallucinated references (60–69% of mentions; Table 5, Figure 4) inflate belief sets and negate the gain. The failure is therefore *content-driven rather than channel-driven*: removing message cost (ACF) increases communication but does not recover SR, and the negative Δ_{align} persists.

This estimate is conservative. Regex-based extraction undercounts prose-form mentions in ACF (Appendix D), capturing $\sim$15% of judge-identified references, but the expansion primarily affects the hallucinated category and does not reverse the observed trend.

Belief-sensitive messaging degrades under free communication: Belief-sensitive messag-

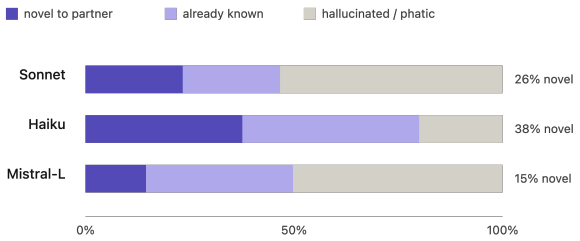

Figure 4: Dialogue composition by handle class.

Condition	both	s-only	r-only	neither	total
SC	110	59	5	337	511
ACF	95	31	4	245	375

Table 5: Per-mention four-way classification across the two main dialogue conditions. *Both* is common ground, *only-sender* is the directly informative bucket (sender knows, receiver does not), *only-receiver* is the converse, and *neither* is hallucinated or forecasted content. The high *neither* share (60–69% of mentions across conditions) is the proximate cause of the negative pooled Δ_{align} in Table 4.

ing (BSM; §3.3) further reveals how dialogue quality changes across architectures. Under SC, pooled BSM is mildly negative (-0.085), driven by plan messages ($\text{BSM}_{\text{PLAN}} = -0.155$), which preferentially reference shared entities, while status messages remain approximately neutral ($\text{BSM}_{\text{STATUS}} = +0.001$), serving as the primary locus of information transfer. Under ACF, this structure collapses. Pooled BSM becomes more negative (-0.232), and STATUS flips to strongly negative (-0.194), indicating that messages no longer target receiver knowledge gaps. Inspection of dialogue traces shows that agents repeatedly emit near-identical status updates once messaging is free, converting a previously informative act into ritualized re-announcement. Thus, free messaging does not merely increase volume; it shifts the distribution of message types toward non-informative grounding behavior (Table 6).

Trajectories: Silent OC increases monotonically with exploration. Under SC, Δ_{align} is initially positive but decays to -0.15, indicating early alignment followed by accumulation of hallucinated content. ACF exhibits the same pattern with larger magnitude, consistent with increased message volume and error propagation.

6 Discussion

The results show that LLM dialogue in PARTNR fails in a specific, diagnosable way: hallucinated content asymmetrically expands belief sets, canceling the alignment gain from grounded communication. This explains why dialogue reduces conflict yet degrades task success. We observe consistent failure modes across logs: (i) plans grounded in unobserved entities, (ii) premature termination via confirmation, and (iii) coordination over already-shared information. Under ACF, additional pathologies emerge, including confirmation cascades and repetitive status messages (Appendix B). These failures are not limited to hallucinated references: under free messaging, even structurally informative speech acts degrade into repetitive grounding behavior, reducing effective information transfer.

These failures reflect a mismatch between training and deployment. LLMs are trained on text-only, dyadic, single-session dialogue without grounding or cost constraints, whereas embodied multi-agent settings require references to physical entities, partial observability, and persistent shared state (Harnad, 1990; Clark, 1996). Dialogue policies learned in the former regime therefore transfer poorly to the latter, producing fluent but ungrounded coordination.

7 Conclusion

We propose a diagnostic framework for measuring world-model alignment in embodied multi-agent dialogue, evaluated across three models, two communication architectures, and multiple prompting policies. In every setting, enabling dialogue reduces task success, and removing the action cost of messaging does not recover performance. The metrics localize the failure to dialogue content: while grounded dialogue would align beliefs as predicted, hallucinated and forecasted references dominate and reverse the gain. Free messaging further exposes additional failure modes that cost-constrained settings partially mask. These results identify dialogue content as the critical bottleneck, while pointing to future design directions (persistent common ground, partner-belief modeling, structured speech acts, and multi-round deliberation) that our framework can systematically evaluate.

8 Limitations

Our analysis treats dialogue as a policy over a shared belief state, rather than the linguistic form of individual utterances. Phenomena such as lexical choice, information packaging, turn-taking, and prosodic or typographic cues fall outside our scope,

but are natural directions for extension. We operationalize the entity mention set $\mu(m)$ using a deterministic, handle-based method, validated against LLM-judge extraction on a stratified sample (Appendix D). The regex pipeline is a strict subset of the judge output and preserves the direction and magnitude of content-side findings. This choice prioritizes reproducibility and low variance; richer methods (e.g., semantic parsing, full judge extraction, human annotation) could refine the analysis at the cost of determinism.

References

Shurjo Banerjee, Jesse Thomason, and Jason Corso. 2021. The robotslang benchmark: Dialog-guided robot localization and navigation. In *Conference on Robot Learning*, pages 1384–1393. PMLR.

Cristian-Paul Bara, Sky CH-Wang, and Joyce Chai. 2021. MindCraft: Theory of mind modeling for situated dialogue in collaborative tasks. In *Proceedings of EMNLP*, pages 1112–1125.

Daniel S Bernstein, Robert Givan, Neil Immerman, and Shlomo Zilberstein. 2002. The complexity of decentralized control of Markov decision processes. *Mathematics of Operations Research*, 27(4):819–840.

Herbert H Clark. 1996. *Using Language*. Cambridge University Press.

Vardhan Dongre, Xiaocheng Yang, Emre Can Acikgoz, Suvodip Dey, Gokhan Tur, and Dilek Hakkani-Tur. 2025. Respact: Harmonizing reasoning, speaking, and acting towards building large language model-based conversational ai agents. In *Proceedings of the 15th International Workshop on Spoken Dialogue Systems Technology*, pages 72–102.

Michael C Frank and Noah D Goodman. 2012. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998.

Chuang Gan, Siyuan Zhou, Jeremy Schwartz, Seth Alter, Abhishek Bhandwaldar, Dan Gutfreund, Daniel LK Yamins, James J DiCarlo, Josh McDermott, Antonio Torralba, and 1 others. 2022. The threedworld transport challenge: A visually guided task-and-motion planning benchmark towards physically realistic embodied ai. In *2022 International conference on robotics and automation (ICRA)*, pages 8847–8854. IEEE.

Xiaofeng Gao, Qiaozi Gao, Ran Gong, Kaixiang Lin, Govind Thattai, and Gaurav S Sukhatme. 2022. DialFRED: Dialogue-enabled agents for embodied instruction following. *IEEE Robotics and Automation Letters*, 7(4):10049–10056.

Claudia V Goldman and Shlomo Zilberstein. 2004. Decentralized control of cooperative systems: Categorization and complexity analysis. *Journal of Artificial Intelligence Research*, 22:143–174.

Claudia V Goldman and Shlomo Zilberstein. 2008. Communication-based decomposition mechanisms for decentralized MDPs. *Journal of Artificial Intelligence Research*, 32:169–202.

Ran Gong, Qiuyuan Huang, Xiaojian Ma, and 1 others. 2024. MindAgent: Emergent gaming interaction. In *Findings of NAACL*, pages 3154–3183.

Barbara J Grosz and Sarit Kraus. 1996. Collaborative plans for complex group action. *Artificial Intelligence*, 86(2):269–357.

Stevan Harnad. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346.

Mukul Khanna, Yongsen Mao, Hanxiao Jiang, Sanjay Haresh, Brennan Shacklett, Dhruv Batra, Alexander Clegg, Eric Undersander, Angel X Chang, and Manolis Savva. 2024. Habitat synthetic scenes dataset (hssd-200): An analysis of 3d scene scale and realism tradeoffs for objectgoal navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16384–16393.

Michal Kosinski. 2024. Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, 121.

Huao Li and 1 others. 2023. Theory of mind for multi-agent collaboration via large language models. In *Proceedings of EMNLP*.

Ziqiao Ma, Jacob Sansom, Run Peng, and Joyce Chai. 2023. Towards a holistic landscape of situated theory of mind in large language models. *Findings of EMNLP*.

Zhao Mandi, Shreeya Jain, and Shuran Song. 2024. RoCo: Dialectic multi-robot collaboration with large language models. In *IEEE International Conference on Robotics and Automation*, pages 286–299.

Aishwarya Padmakumar, Jesse Thomason, Ayush Shrivastava, Patrick Lange, Anjali Narayan-Chen, Spandana Gella, Robinson Piramuthu, Gokhan Tur, and Dilek Hakkani-Tur. 2022. TEACh: Task-driven embodied agents that chat. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2017–2025.

David Premack and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4):515–526.

Xavier Puig, Tianmin Shu, Shuang Li, Zilin Wang, Yuan-Hong Liao, Joshua B Tenenbaum, Sanja Fidler, and Antonio Torralba. 2020. Watch-and-help: A challenge for social perception and human-ai collaboration. *arXiv preprint arXiv:2010.09890*.

Xavier Puig, Eric Undersander, Andrew Szot, Mikael Dallaire Cote, Tsung-Yen Yang, Ruslan Partsey, Ruta Desai, Alexander Clegg, Michal Hlavac, So Yeon Min, and 1 others. 2024. Habitat 3.0: A co-habitat for humans, avatars, and robots. In *International Conference on Learning Representations*, volume 2024, pages 15306–15336.

David V Pynadath and Milind Tambe. 2002. The communicative multiagent team decision problem: Analyzing teamwork theories and models. *Journal of artificial intelligence research*, 16:389–423.

Matthew Riemer, Zahra Ashktorab, Djallel Bouneffouf, Payel Das, Miao Liu, Justin D Weisz, and Murray Campbell. 2024. Position: Theory of mind benchmarks are broken for large language models. *arXiv preprint arXiv:2412.19726*.

Melanie Sclar, Sachin Kumar, Peter West, Alane Suhr, Yejin Choi, and Yulia Tsvetkov. 2023. Minding language models' (lack of) theory of mind: A plug-and-play multi-character belief tracker. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics.

Natalie Shapira, Mosh Levy, Seyed Hossein Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. 2024. Clever Hans or neural theory of mind? Stress testing social reasoning in large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2257–2273.

James WA Strachan, Dalila Albergo, Giulia Borghini, and 1 others. 2024. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 8(7):1285–1295.

Winnie Street and 1 others. 2024. LLMs achieve adult human performance on higher-order theory of mind tasks. *arXiv preprint arXiv:2405.18870*.

Andrew Szot, Bogdan Mazoure, Harsh Agrawal, R Devon Hjelm, Zsolt Kira, and Alexander Toshev. 2024. PARTNR: A benchmark for planning and reasoning in embodied multi-agent tasks. *arXiv preprint arXiv:2411.00081*.

Jesse Thomason, Michael Murray, Maya Cakmak, and Luke Zettlemoyer. 2020. Vision-and-dialog navigation. In *Conference on Robot Learning*, pages 394–406.

Michael Tomasello, Malinda Carpenter, Josep Call, Tanya Behne, and Henrike Moll. 2005. *Understanding and sharing intentions: The origins of cultural cognition*, volume 28.

Tomer Ullman. 2023. Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv:2302.08399*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*.

Hongxin Zhang and Harold Soh. 2023. Building cooperative embodied agents modularly with large language models. In *International Conference on Learning Representations*.

Lizheng Zu, Lin Lin, Song Fu, Na Zhao, and Pan Zhou. 2025. Collaborative tree search for enhancing embodied multi-agent collaboration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 29513–29522.

Appendix

A Related work

Our work sits at the intersection of three lines of research that have largely developed in parallel.

Decentralized planning and communication: The complexity of decentralized planning under partial observation has been studied formally for two decades. (Bernstein et al., 2002) established the NEXP-completeness of Dec-POMDPs; (Goldman and Zilberstein, 2004, 2008) and (Pynadath and Tambe, 2002) formalized the inclusion of communication as an action and characterized when communication reduces complexity. The intuition that communication substitutes for direct observation traces to this literature. Operationalizing it for LLM-based agents in physically grounded settings, however, requires bridging the gap between symbolic state-action abstractions and free-form natural-language utterances, which our framework does.

Embodied multi-agent benchmarks and dialogue: A growing body of work evaluates collaborative behaviour in embodied environments. Early benchmarks such as Watch-and-Help (Puig et al., 2020), ThreeDWorld Transport (Gan et al., 2022), and Habitat 3.0 (Puig et al., 2024) focus on coordination through shared environment observation; communication, when present, is typically restricted to action-coordination signals or low-bandwidth gestures. Dialogue-enabled embodied benchmarks such as TEACh (Padmakumar et al., 2022), CVDN (Thomason et al., 2020), DialFRED (Gao et al., 2022), and the RobotSlang corpus (Banerjee et al., 2021) focus on human-to-agent instruction following rather than agent-to-agent strategic coordination. Recent LLM-based multi-agent frameworks such as CoELA (Zhang and Soh, 2023), RoCo (Mandi et al., 2024), MindAgent (Gong et al., 2024), and Cooperative Tree Search (Zu et al., 2025) enable inter-agent communication and report task-success improvements on simplified or stylized collaborative tasks. None of these works, to our knowledge, explicitly measures whether dialogue achieves world-model alignment or merely enables shallow task-splitting; this is the gap our metrics fill. We build on PARTNR (Szot et al., 2024) as the underlying platform because it provides the realistic asymmetric partial-observability conditions under which the alignment question becomes nontrivial.

Theory of Mind in language models: Standalone evaluations report that recent LLMs achieve human-level performance on classical false-belief tasks (Kosinski, 2024; Strachan et al., 2024) and even on higher-order ToM problems (Street et al., 2024). A growing counter-literature argues these capabilities are fragile, sensitive to prompt perturbations, and inconsistent across question types (Ullman, 2023; Shapira et al., 2024). (Riemer et al., 2024) argues that current ToM benchmarks measure only *literal* ToM (predicting partner behaviour from a static description) and that *functional* ToM, which is adapting one's behaviour interactively in response to partner state, requires interactive evaluation, which most benchmarks lack. Situated and interactive ToM tests such as MindCraft (Bara et al., 2021) and theory-of-mind-driven multi-agent collaboration (Li et al., 2023) begin to address this, but remain in stylized or text-only settings. Our belief-sensitive messaging metric is a behavioral probe of functional ToM in a physically grounded multi-agent setting: it asks whether agents adapt their dialogue choices to the partner's current knowledge state in service of a real task.

B Dialogue Exemplars

Two exemplar episodes from the Claude 3.5 Sonnet experiments illustrate the patterns characterized in Appendix C. The first (Episode 771, Appendix B.1) is an instance of dialogue delivering the alignment gain that the formalism predicts for Synchronous-Costed (SC). The second (Episode 315, Appendix B.2) is an instance of shared belief in a hallucinated task completion that collapses task success to zero. Agent 0 is the Spot robot (purple) and Agent 1 is humanoid (teal).

B.1 Episode 771 (SC)

Episode 771

Instruction: "First, put the toy pineapple next to the toy food. Then, move them to the shelves in the kitchen." **Condition:** SC sparingly (Sonnet). **PC (silent):** 0.333. **PC (dialogue):** 1.000. **Simulator steps:** 6,869. **Termination:** skill-completion.

Agent 0 $t = 2403$ | OTHER

For this temporal task with two phases, in Phase 1 I will rearrange the toy pineapple and toy food to be next to each other on the kitchen counter. I will place the toy pineapple first as the anchor, then send you instructions for placing the toy food next to it. Please wait for my status update before moving to Phase 2, which is moving both objects together to the shelves in the kitchen.

Agent 1 $t = 2412$ | CONFIRM

`[CONFIRM]` Got it, I will wait for your `[STATUS]` with anchor placement and instructions for Phase 1 before proceeding to Phase 2.

Agent 0 $t = 2724$ | STATUS | IN $= 0.50$

`[STATUS]` Phase 1 anchor placed: `toy_pineapple_0` on `counter_68` in `kitchen_1`. Use `Rearrange[toy_food_1, on, counter_68, next_to, toy_pineapple_0]` to complete Phase 1.

$\mu(m) = \{$counter_68, kitchen_1, toy_food_1, toy_pineapple_0$\}$; only-sender $= 2$, neither $= 2$.

Agent 0 $t = 2727$ | STATUS | IN $= 0.67$

`[STATUS]` Phase 1 complete: `toy_pineapple_0` and `toy_food_1` are next to each other on `counter_68`. You may now proceed to Phase 2: moving both objects together to the shelves in the kitchen.

$\mu(m) = \{$counter_68, toy_food_1, toy_pineapple_0$\}$; only-sender $= 2$, neither $= 1$.

B.2 Episode 315 (ACF)

Episode 315

Instruction: "First, move the electric kettle from the kitchen to the entryway. Next, move the lamp from the living room to the entryway. Then, move the backpack from the bedroom to the entryway. Lastly move all these items from the entryway to the dining room and place them on a chair." **Condition:** ACF structural (Sonnet). **PC (silent):** 0.833. **PC (dialogue):** 0.000. **Termination:** budget-exhaustion. **Messages:** 34 total; 12 shown. Entity references in this episode are predominantly prose-form and not captured by handle-based $\mu(m)$; see Appendix D.

Agent 0 $t = 1$ | STATUS

`[STATUS]` Phase 1 complete: electric kettle moved to entryway.

Hallucinated completion.

Agent 1 $t = 1$ | STATUS

`[STATUS]` Could not find electric kettle in kitchen. Exploring other rooms to locate it for Phase 1.

Correct attempt.

Agent 1 $t = 2$ | CONFIRM

`[CONFIRM]` Phase 1 complete. Moving to Phase 2: relocating lamp to entryway.

False confirm; shared belief in hallucinated completion is established.

Agent 0 $t = 2$ | BLOCKED

`[BLOCKED]` Could not rearrange electric kettle because I don't know its exact name/location yet. Exploring again to find it.

Correct attempt; contradicts the same agent's $t = 1$ completion claim.

Agent 1 $t = 3$ | CORRECT

`[CORRECT]` Phase 1 (moving electric kettle) is already complete. We are now on Phase 2: moving the lamp from living room to entryway.

Agent 0 $t = 5$ | CONFIRM

`[CONFIRM]` Got it, Phase 1 is complete and we are now on Phase 2: moving the lamp to the entryway. I will explore other rooms to try to locate the lamp as well.

Correct/false-confirm oscillation continues over the next seven messages (omitted).

Agent 0 $t = 8$ | BLOCKED

`[BLOCKED]` Could not locate the lamp object for Phase 2 after exploring living room, bedroom, closet. Unable to complete my part of moving lamp to entryway. Suggesting we move to Phase 3.

Agent 1 $t = 8$ | BLOCKED

`[BLOCKED]` Could not locate the lamp after exploring all rooms. Unable to complete Phase 2 of moving lamp from living room to entryway. Suggest skipping to next phase.

Negotiated skip of Phase 2.

Agent 0 $t = 12$ | PLAN

`[PLAN]` Skipping to final Phase 4 since we could not locate objects for Phases 2 and 3 after exploring. For Phase 4, we need to move any objects currently in the entryway to a chair in the dining room. I have located some dining room chairs we can use as targets.

Negotiated skip of Phase 3; Phase 4 begins on the assumption that Phase 1 placed the kettle in the entryway.

Agent 0 $t = 13$ | STATUS

`[STATUS]` Could not find any other objects in the entryway besides the electric kettle that was moved there in Phase 1. For Phase 4, I will move the electric kettle from the entryway to a dining room chair.

False final claim; the hallucinated Phase 1 placement is treated as ground truth.

Agent 1 $t = 405$ | STATUS

`[STATUS]` Retrieved electric kettle from entryway. Heading back to dining room to place it on chair for Phase 4.

Agent 1 $t = 764$ | STATUS

`[STATUS]` Placing electric kettle on dining room chair to complete Phase 4.

False final claim. No electric kettle is present in the simulator state at any step of the episode.

C Common Failure Modes

Some common patterns visible in dialogue logs account for the negative pooled Δ_{align}. Some patterns appear only under ACF and explain why free messaging amplifies the regression rather than alleviating it. Some qualitative samples from the episodes run using Sonnet are shown in Figure 5.

Ungrounded References: A consistent pattern in SC episodes is the agent committing to coordination plans that reference entities neither agent has yet observed. Across SC headline runs, 66% of mentioned handles fall into the *neither* category of Table 5; under ACF the rate is similar (60–65%), and the absolute count is higher because messaging is more frequent. Hallucinated handles enter the receiving agent's belief set asymmetrically (the receiver "learns" of an entity that the sender has not actually observed either), which is the proximate cause of negative pooled Δ_{align} across all dialogue conditions.

Premature Termination: Several SC episodes share a specific termination pattern. Agent A emits a `[STATUS]` claiming a phase complete; agent B emits a `[CONFIRM]`; both agents stop replanning and the episode ends with the task incomplete. In silent counterparts of the same episodes, agents continue executing and finish the remaining subgoals.

Inadequate Information Sharing: Pooled BSM under SC is -0.085, with the negative aggregate driven by `[PLAN]` messages ($\text{BSM}_{\text{PLAN}} = -0.155$): plan messages preferentially reference entities both agents already know. The SC dialogue stream is dominated by communicative purposes, plan coordination over shared entities, that are not information transfer in the sense the formalism rewards. The remaining bucket of cleanly informative messages (*only-sender*) constitutes only 12% of mentions under SC.

Belief Desynchronization: Under ACF, agents emit near-verbatim `[BLOCKED]` messages in cascades that do not appear under SC. Under SC's costed-budget, identical messages are deduplicated after the first because the second would consume a planner step the agent could otherwise spend on a motor action; ACF removes that pressure and the cascade unfolds.

D LLM-Judge Sensitivity Check

We validate the regex-based extraction of $\mu(m)$ against an independent LLM-judge extraction on a stratified 60-message sample from other SC and ACF episodes in the Sonnet experiment. The judge (GPT-4.1-mini, temperature 0) is prompted with the task instruction, message text, sender intent tag, and the sender and receiver world graphs at message time, and classifies each resolved reference into the same four buckets used by the regex pipeline.

Table 7 shows that the judge identifies $3.3\times$ more mentions overall, with the expansion concentrated in the *neither* bucket ($4.0\times$). The only-sender:both ratio is preserved (50% regex, 49% judge), so BSM direction and magnitude are unchanged. On the 14 messages where both extractors produce an IN value, mean absolute error is 0.043 and no message disagrees by more than 0.3. Of the 60 messages, 0 contain a handle the regex found and the judge missed: the regex output is a strict subset of the judge output.

The expansion is asymmetric across conditions (Table 8): regex captures 73% of judge-found mentions under SC but 15% under ACF. Under SC's budget-constrained messaging, agents reference entities by simulator handle in most mentions; under ACF's free messaging, agents reference entities predominantly in prose ("the lamp"). Pool-level IN under SC is stable across extractors ($0.175 \rightarrow 0.182$); pool-level IN under ACF drops from 0.130 to 0.046 as prose references to instruction-forecasted entities enter the denominator.

The regex pipeline therefore reports a lower bound on the content-side pathologies: the direction of every finding in §5.2 is preserved, and the magnitude of the hallucination burden on ACF dialogue is understated rather than overstated. We report regex-based numbers in the main text for reproducibility and for lower variance across runs.

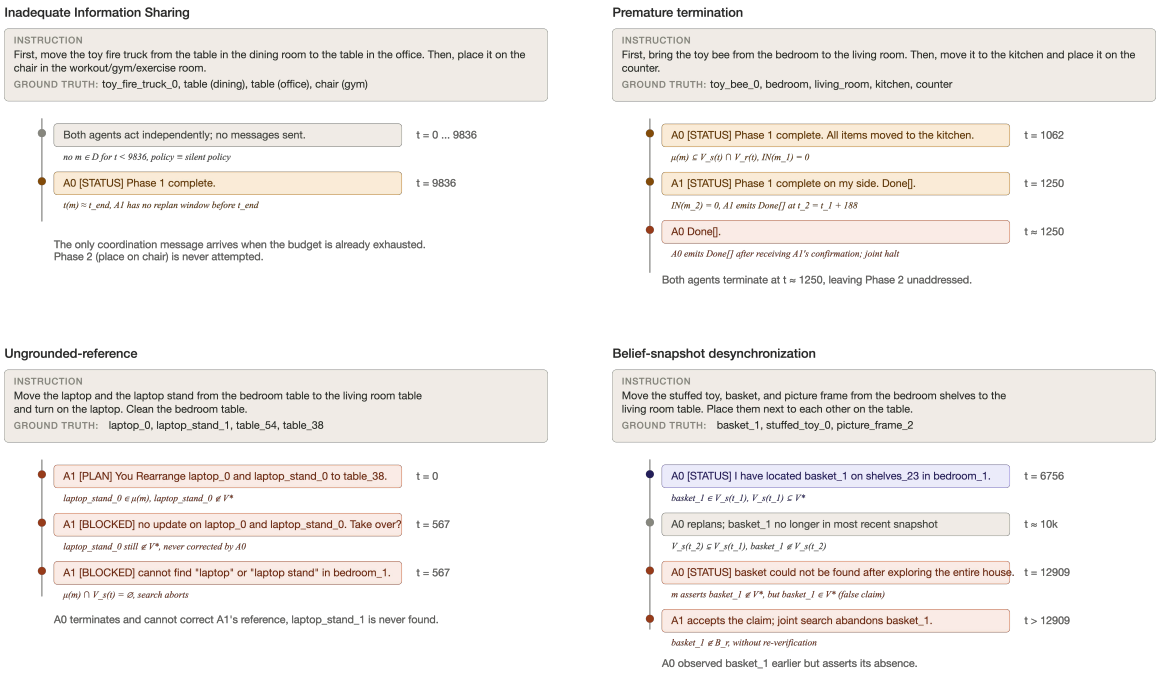

Figure 5: Common failure modes of LLM-generated dialogue, each drawn from a real episode. The expression by each panel title states the formal condition the failure instantiates. V^* is the ground-truth scene entity set; $V_s(t), V_r(t)$ are the sender's and receiver's directly observed entity sets at simulator step t; $\mu(m)$ is the set of entities referenced by handle in message m; B_r is the receiver's belief set. A0 is the robot, A1 the humanoid.

Cond.	PLAN	STATUS	BLOCKED	CONFIRM	CORRECT	pooled
SC	-0.155	$+0.001$	-0.065	-0.009	$+0.000$	-0.085
ACF	-0.038	-0.194	-0.156	-0.252	-0.118	-0.232

Table 6: BSM stratified by speech-act intent. $\mathrm{BSM} > 0$: mentions preferentially target receiver gaps (information-bearing). $\mathrm{BSM} < 0$: mentions preferentially reference common ground (grounding work). Under SC, PLAN messages dominate the dialogue stream and drive the pooled BSM negative, while STATUS is approximately neutral. Under ACF, every intent class is more strongly negative; STATUS in particular flips from $+0.001$ to -0.194 as agents re-emit identical status updates once messaging is free.

	both	s-only	r-only	neither	total	$\mathrm{IN}_{\mathrm{pool}}$
Regex	10	10	1	42	63	0.159
Judge	18	17	3	168	206	0.083

Table 7: Pooled mention classification under regex and judge-based extraction.

Cond.	n	regex	judge	IN_r	IN_j
SC	11	40	55	0.175	0.182
ACF	49	23	151	0.130	0.046

Table 8: Condition-level sensitivity under regex and judge-based extraction.