LLMs and their Limited Theory of Mind: Evaluating Mental State Annotations in Situated Dialogue

Katharine Kowalyshyn and **Matthias Scheutz**
Department of Computer Science
Tufts University
{Katharine.Kowalyshyn, Matthias.Scheutz}@tufts.edu

Abstract

Effective human teams excel at maintaining a consistent *shared mental model* (SMM) that reflects the shared understanding of individual team members about the task and what remains to be done. We present a novel, two-step framework that leverages large language models (LLMs) both as (1) generators of mental model traces from team dialogues to track the team's SMM and (2) automated detectors of discrepancies between inferred and ground truth mental model traces. We define an SMM coherence evaluation framework for this use case and apply it to six dialogues in a previously published team corpus, ultimately producing a dataset of human and LLM SMM mental model traces, a reproducible evaluation framework for SMM coherence, and an empirical assessment of LLM-based discrepancy detection. Our results reveal that while LLMs exhibit apparent coherence on straightforward natural-language trace generation tasks, they systematically err in scenarios requiring spatial reasoning or disambiguation of transcription-level disfluencies.

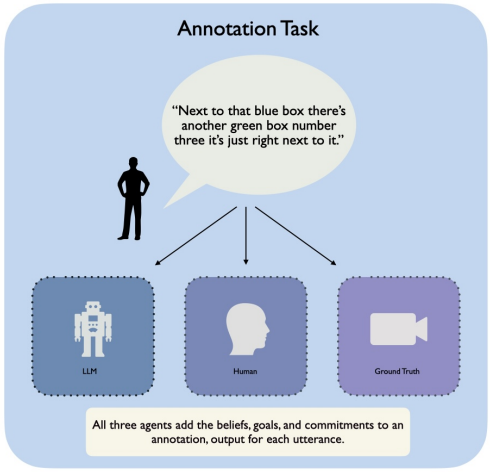


Figure 1: Example of the first stage of our evaluation pipeline. A snippet of situated team dialogue is independently processed by an LLM, a naive human, and a video-informed human with access to the environment, each producing a mental model trace of belief, commitment, and goal states for each agent. The audio-only traces (LLM and naive human) are later compared against the ground truth trace by an LLM judge to assess ToM-like reasoning capabilities.

1 Introduction

Successful team performance often hinges not only on what is said, but on what is understood, including the latent belief states, goals, and commitments shared among team members. In tasks which require coordination through spoken communication, situated dialogue, embedded in real-time decision-making and environmental context, becomes the primary conduit for this shared understanding called *shared mental model* (e.g., (Scheutz et al., 2017)). For humans, the subtext behind task-based communications (i.e., the assumptions, intentions, and knowledge of others) is often grasped implicitly and the question we ask is whether large language models (LLMs) could do the same. More precisely, can LLMs accurately infer and thus annotate internal belief states of team members based on dialogue interactions and detect discrepancies in their mental representations?

In this paper, we treat situated dialogue as a proxy task for assessing *Theory of Mind* (ToM) capabilities in current large language models, focusing on team coherence in situated dialogue. We do so by utilizing the *Cooperative Remote Search Task* (CReST) corpus (Eberhard et al., 2010) which consists of situated task-based interactions between a local *searcher* and remote *director*. As part of our experimental setup, we evaluate three independent LLMs, OpenAI's o3-mini, Anthropic's Claude Sonnet 4, and Google's Gemma, alongside a pair of naive humans and another pair of humans with access to the ground truth, to each generate mental model traces for six CReST dialogues for which video recordings are available to

Proceedings of the 27th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 738–756
August 2–5, 2026.

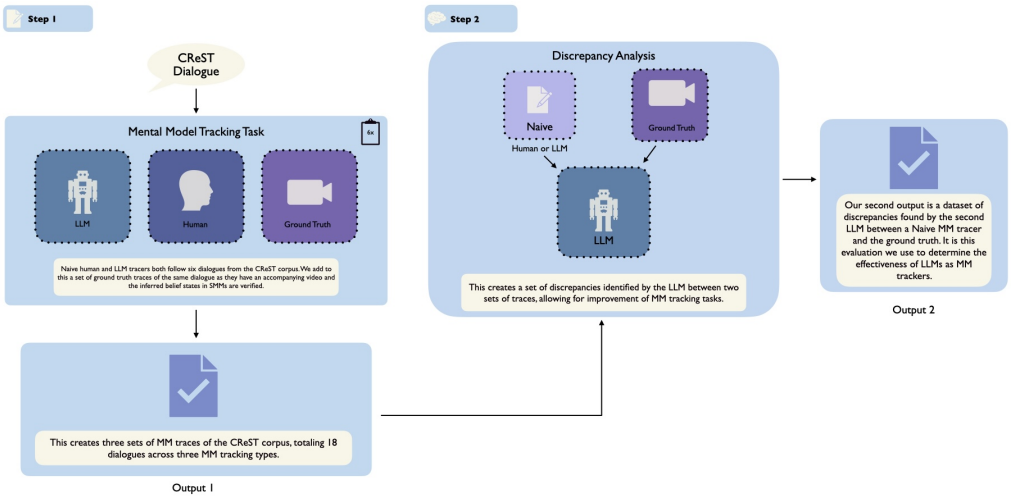


Figure 2: An overview of our two-step process used to trace and find discrepancies between SMMs using LLMs. We are able to create two outputs: (1) a dataset of ground truth, human, and LLM mental model traces and (2) a dataset of discrepancies between ground truth traces and human/LLM mental model traces.

establish ground truth traces [1]. We then employed a separate "LLM judge" to compare the audio-only traces (both human and LLM) against the ground truth traces, which we call "discrepancy analysis". Additionally, we outline a discrepancy evaluation framework which can apply to many team-based coordination tasks, extending the utility of this paper beyond the empirical results we provide.

This two-step approach (as seen in Figure 2) yields a dataset of mental model traces, a reproducible evaluation pipeline defined by fixed prompts, tracing procedures, discrepancy metrics, and an empirical assessment of LLM discrepancy detection. While our primary goal is to assess LLMs' ability to trace belief states in team dialogue, our findings also underscore a broader point: LLMs simulate limited but measurable ToM-like behaviors in situated contexts. To our knowledge, this is the first study to use belief-state traces in real-world dialogue as a lens for evaluating ToM in LLMs. As such, we propose situated dialogue trace generation as a viable proxy task for probing ToM capabilities in AI systems. The remainder of this paper is structured as follows: Section 2 reviews related work on LLMs as generators of mental models and team coherence. Section 3 describes the CReST corpus. Section 4 introduces our discrepancy framework. Section 5 presents results, and Section 6 discusses implications and limitations.

[1] 3 LLMs × 6 dialogues + 1 naive human pair × 6 dialogues + 1 ground truth human pair × 6 dialogues = 30 total trace sets. The 24 comparisons derive from comparing each of the 4 non-ground-truth trace generators across 6 dialogues.

2 Related Work

LLMs have shown impressive capabilities for natural language understanding and for performing novel tasks such as interpreting natural language expressions and mapping them into new formalisms, from programs, to rules, to logics (Pan et al., 2023; Manas et al., 2024; Zhou et al., 2024). While recent work has explored ToM in LLMs, many existing evaluations focus on static, decontextualized tasks such as multiple-choice questions, short story comprehension, or isolated inference problems. These tasks often test factual or scripted understanding rather than the dynamic, real-time reasoning that underlies ToM in human communication (Chen et al., 2024; Kosinski, 2024; Strachan et al., 2024; Gandhi et al., 2023). In contrast, situated dialogue involves ongoing belief tracking, contextual inference, and implicit coordination, which are abilities more reflective of how ToM functions in natural human interaction. Recently there has even been work negating the idea that so-called thinking/reasoning models can reason at all (Shojaee*† et al., 2025). This highlights a critical gap in the current literature: few, if any, ToM evaluations challenge models to reason about mental states as they unfold in interactive, grounded team contexts. Here we briefly review recent work on using LLMs as for tracking human mental states in task-based dialogues.

2.1 LLMs for Mental State Inference

LLMs have rapidly become integral to modern corpus construction and annotation, spanning direct label suggestion, iterative human–LLM editing, and

conversational scaffolding (Yu et al., 2024; Zhu et al., 2025; Wu et al., 2024; Zhao et al., 2024; Weissweiler et al., 2024; Pangakis and Wolken, 2024). Together, these approaches reduce human workload while improving annotation consistency and scalability across a wide range of NLP tasks. Still, no prior work combines LLM-generated mental model traces with the inference of mental states in team dialogue.

2.2 LLMs for Team Coherence

Recent work in multi-party conversation tracking, role-playing, and peer agents all include LLMs as a main character in the task (Addlesee et al., 2023; Kong et al., 2024; Liu et al., 2024a; Palmer et al.), but LLM reasoning abilities have repeatedly been questioned despite their benchmark performance (Stechly et al., 2024; Mondorf and Plank, 2024; Bertolazzi et al., 2024). While some hypothesize that ToM reasoning has already emerged as a side effect of training data (Ma et al., 2023; Kosinski, 2024; Kim et al., 2023), others argue that, overall, experiments where LLMs have performed well on ToM tasks are not enough evidence for the emergence of cognitive abilities akin to humans (Sarıtaş et al., 2025; Ullman, 2023).

In the teaming context, the notion of a *shared mental model* began alongside cognitive science's inception (Johnson-Laird, 1980), and various psychological studies have been performed to understand team performance in the past, including to generate natural dialogue and evaluate shared mental models (Mathieu et al., 2000; Gervits et al., 2016; Jonker et al., 2010). Extending this work to human-AI teams research has included evaluations of SMMs and their impact on team performance (Andrews et al., 2023; Schelble et al., 2022; Sarkar et al., 2025a), including human-robot team performance, establishing that task tracking through SMMs enhances team performance (Scheutz et al., 2024; Edgar et al., 2024). Recent team coherence research also includes the creation of collaborative teams of LLMs and LLM understanding of utterances with uncertainty (Liu et al., 2024b; Paige et al., 2024). Adjacent work on dialogue state tracking and commitment recognition addresses related inference problems but typically focuses on surface task states rather than the full mental model of each agent, including second-order beliefs (Sarkar et al., 2025a).

Despite rich work on human SMMs and on LLM mental state inference in isolation, no prior work evaluates LLMs' ability to infer mental states in team-based dialogues. Overall, given that SMMs are useful for human teams, that humans are able to build and update their mental models of other team members through dialogues, and that there is an increasing interest in mixed-initiative human-AI teams, the question addressed in this paper is timely – to our knowledge there is no other work that investigates the ability of LLMs to make mental state inferences from task-based dialogues and compares them to human trace generation performance.

3 Dataset: The CReST Corpus

To investigate the ability of LLMs for dialogue-based inference about human mental states, we utilized the "Cooperative Remote Search Task" (CReST) corpus which is comprised of task-based dialogues from 23 human dyad teams and was originally created to fill the void of a lack of naturalistic dialogue in team-based settings (Eberhard et al., 2010). Each group consisted of one person designated to be a searcher who was present in an environment, while the other person, the director, was not physically present but was given a map of the space which was intentionally incomplete (full map is shown in Appendix B). The pair communicated via audio channel and were given a series of tasks to complete. Tasks included locating boxes in the environment, labeling boxes on the map, and changing the locations of blocks which were found within specifically-colored boxes.

The resulting corpus contains 40,083 words in 5,872 sentences between all twenty-three pairs (Kübler et al., 2012; Eberhard et al., 2010; Gervits et al., 2016). Crucial to this paper, this corpus highlights disfluencies and typical speech patterns, which allowed us to conduct a more natural simulation of SMM tracking in real time situated dialogues. Here, we use the transcribed text from six dialogues to generate mental model traces for which additional video recordings were available from the searcher's headcam (Video image in Appendix B), which allowed us to determine the ground truth of where the searcher was located and what the searcher was seeing at all times during the task-based interactions (for the stationary director this was not necessary). We also generate traces from the same dialogues using naive annotators sans access to the videos, to explore the agreement rate between human annotators and LLMs[2] The

[2]Please see the Appendix for examples and prompts.

following sections outline the mental model tracking procedure for both naive human and LLM, and ground truth inference-based traces.

3.1 Mental Model Tracking Procedure

All of the agents' mental states were first traced by two humans across all dialogues. Throughout this section, we use "annotator" to refer specifically to the human workers hired to label the dialogues; the outputs they produce are mental model traces in the framework described in Section 4. These human annotators were undergraduate students hired and compensated at an hourly rate of $17 in accordance with university policy. We hired two annotators to first naively annotate (without ground truth videos) the dialogues for mental states independently. Annotations were performed in a free-form manner without any templates, but after instructions given by the experimenter with examples of how to track task-relevant updates about locations, goals, etc. Then, the two annotators sat together and resolved any disagreements, creating a dataset of agreed upon naive annotations with an agreement rate of 100% post-adjudication. Following this, we completed the same procedure but with the ground truth videos. See Figure 3 for the annotation format. In total, we had six dialogues consisting of **1,142** utterances and a resulting two-part dataset with ground truth and naive annotations from humans (see Table 1 below for more detail).

Dialogue	Number of Utterances
1	173
2	147
3	225
4	162
5	241
6	194
Total	**1,142**

Table 1: Total utterances for each dialogue in our dataset created from the CReST corpus. This includes all of the dialogues from the corpus which have a corresponding video, which is necessary in order to establish ground truth mental state traces.

We generated our dataset (Step 1 in Figure 2) by having three groups of trace generators label the same six dialogues for the searcher's and director's beliefs, goals, and commitments. The first set of trace generators was a series of naive LLMs which were naive as they were only given limited information about the environment: prompts detailed task

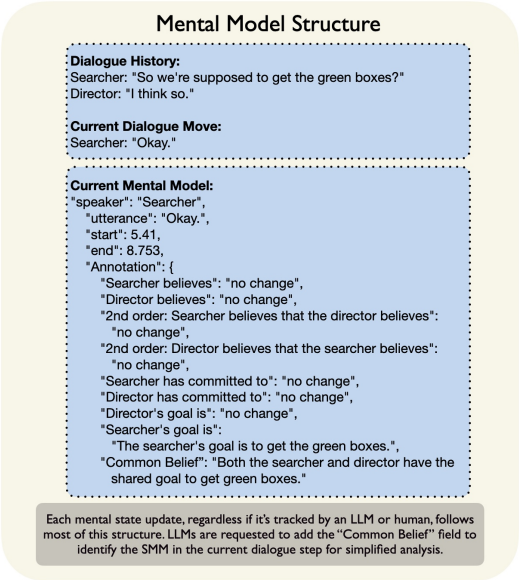


Figure 3: Using an example from Dialogue 1, we show the trace structure used by LLMs to identify shared mental models and the common beliefs held by agents. Our dataset of traces includes a mental state update of this structure for each utterance.

overview (see Appendix A for prompts) but there was no plausible way for the LLM to establish a ground truth understanding of the environment operating on natural language interaction alone; all updates about the state of the environment had to be inferred from the utterances. The second set of mental state trackers were two "naive humans" who similarly only had the language dialogues available and were compensated and completed the adjudication process as outlined above. Figure 3 shows the annotation structure for each utterance, which was employed by both human and LLM mental state tracers.

For the third trace generation condition, we employed two humans to generate mental state traces in the same way, but this time with access to a video which provides ground truth environmental state from the searcher's point-of-view recorded from a headcam as well as the director's screen with the map, and audio between the two. This ground-truth traced mental state set, corroborated by two humans, served as our ground truth comparison for the LLMs' (and the naive humans') mental trace performance.

The key strength of our approach is the use of video recordings to establish ground truth mental model traces. The use of video recordings to establish ground truth traces, and the distinction between audio-only and video-grounded inference, are discussed in detail in Section 4.

Following the creation of our dataset, we were able to move on to Step 2 of Figure 2, the tracing discrepancy analysis.

4 Discrepancy Analysis Framework

There are two separate applications of LLMs in this work: one where an LLM generates mental model traces for the dialogue agents, and a second where a separate LLM judge identifies discrepancies between those traces and the ground truth traces derived from video. For the latter, the question arises how to detect and measure discrepancies between an inferred mental model trace and the ground truth trace. We therefore first introduce our discrepancy framework and metric, then apply it to our dataset.[3]

It is important to distinguish two sources of divergence between an audio-only trace and the ground truth trace:

1. **Trace divergence due to missing perceptual information:** cases where the ground truth trace contains a mental model update that cannot be inferred from the transcript alone, because it derives from a perception or action not verbalized by either agent.
2. **Mental model discrepancies:** cases where the audio-only trace assigns a different belief, goal, or commitment than the ground truth trace even for fields that *could* in principle be inferred from the transcript.

These two sources of divergence are related: trace divergence implies mental model discrepancies. However, the converse does not hold: discrepancies in individual mental model fields may arise from inferential errors rather than missing perceptual information. Our discrepancy framework operates at the level of individual field comparisons between the audio-only and ground truth traces.

4.1 Discrepancy Types

We define a *discrepancy instance* as a mismatch between a single field in an audio-only mental model state and the corresponding field in the ground truth state at the same utterance. Each field is evaluated independently, so discrepancies are counted at the field level. Our framework introduces four types of discrepancies, which we now define using a running example, although other types of discrepancies are possible. Let $M_t^*(x)$ denote the "ground truth" mental model state of person x at utterance t derived from video, and let $M_t(x)$ denote the mental model state for person x at the same utterance derived from audio only. Each mental model consists of a set of propositions over agents and world states.

Consider then the two mental models after utterance t:

$$M_t^*(\text{searcher}) = \{\texttt{pink\_box}(r_3),\ \texttt{at}(\text{searcher}, r_3)\}$$

$$M_t(\text{searcher}) = \{\texttt{pink\_box}(r_3),\ \texttt{at}(\text{searcher}, r_2)\}$$

Here the audio-only trace assigns the searcher to room 2 while the ground truth places them in room 3. This mismatch constitutes a discrepancy instance in the `at(searcher, ·)` field. The four discrepancy types below classify the nature of such mismatches.

4.1.1 Belief Contradictions

A **belief contradiction** occurs when, at the same utterance, the searcher and director hold opposing beliefs about the same world state. Using a similar example as above, a belief contradiction would be:

$$M_t(\text{searcher}) \ni \texttt{at}(\text{green box}, r_3),$$
$$M_t(\text{director}) \ni \texttt{at}(\text{green box}, r_2)$$

These are mutually exclusive propositions about the green box's location, and the co-occurrence across the two traces constitutes a belief contradiction.

4.1.2 Omissions

An **omission** occurs when M_t^* contains a proposition p for a given agent that is absent from M_t, i.e., the audio-only trace simply does not record the corresponding belief or goal update. For example, if the ground truth trace records that the director's goal has shifted to directing the searcher toward corridor 2, but the audio-only trace records *no change* for the director's goal field at that utterance, this constitutes an omission.

4.1.3 Unsupported Beliefs

An **unsupported belief** occurs when M_t contains a proposition p for a given agent that is neither confirmed nor contradicted by M_t^*. It cannot be validated or negated using the ground truth. These often arise when an LLM infers a belief that is linguistically plausible but not grounded in any observable dialogue event or environmental fact.

[3] We will make the entire dataset available on GitHub for external use when the paper is published.

For example, if the audio-only trace records that the searcher's goal is to follow the director's directions at an utterance where the ground truth records no goal update, this is an unsupported belief. Unlike omissions, unsupported beliefs reflect additions to the mental model trace that have no evidential basis, and thus represent a distinct failure mode in which the inferred trace overgenerates.

4.1.4 False Beliefs

A **false belief** discrepancy occurs when an agent's belief disagrees with the ground truth trace, regardless of whether it contradicts the other agent. Suppose M_t contains a proposition p for a given agent and M_t^* contains a proposition $q \neq p$ for that same agent, where p and q are not directly negations of one another (which would constitute a contradiction), but p is nonetheless incorrect given the ground truth. For example:

$$M_t^*(\text{director}) \ni \texttt{pink\_box}(r_3),$$

$$M_t(\text{director}) \ni \texttt{green\_box}(r_3)$$

Here the audio-only trace assigns an incorrect object type to room 3 in the director's mental model state. The ground truth does not assert the negation of $\texttt{green\_box}(r_3)$ directly, but the proposition is nonetheless wrong given the video evidence.

4.2 Severity of Discrepancies

We introduce a severity ranking over the four discrepancy types defined above. This ranking reflects the degree to which each type of discrepancy distorts the inferred mental model trace relative to the ground truth, and is motivated by the downstream impact such distortions would have on team coordination if acted upon.

Belief contradictions and false belief discrepancies represent direct mismatches between the propositions assigned to an agent's mental model state, and therefore most severely distort the inferred trace. An agent acting on a contradicted or false belief would have a fundamentally incorrect understanding of the task state. Unsupported beliefs correspond to spurious additions to the mental model trace that lack evidential grounding; while not directly harmful in isolation, they introduce noise into the inferred trace and could mislead downstream reasoning if acted upon. Omissions are the least severe, as they represent missing updates rather than incorrect ones, though omitting critical state updates can still degrade team coherence when the absent information is task-relevant.

We therefore rank discrepancies in descending order of severity: **Belief Contradictions** $>$ **False Beliefs** $>$ **Unsupported Beliefs** $>$ **Omissions**. In the results reported here, all types are assigned equal weight as a conservative, task-agnostic baseline; differential weighting is left to task-specific adaptation. Appendix E provides per-type, per-utterance breakdowns to support reweighting for future use cases.

To operationalize this ranking, we define a weighted discrepancy score per model–dialogue pair (m, d). Let

- $B_{m,d}$ = number of Belief Contradictions
- $F_{m,d}$ = number of False Beliefs
- $U_{m,d}$ = number of Unsupported Beliefs
- $O_{m,d}$ = number of Omissions
- w_x = weight for respective discrepancy x

Then, the weighted raw discrepancy score is

$$r_{m,d} = w_b B_{m,d} + w_f F_{m,d} + w_u U_{m,d} + w_o O_{m,d}$$

To account for dialogue length, we divide by the number of utterances N_d in dialogue d, resulting in a per-utterance discrepancy score:

$$s_{m,d} = \frac{r_{m,d}}{N_d}$$

Finally, to enable fair comparisons across models and dialogues, we normalize the scores to the range $[0, 1]$:

$$\mathcal{S}_{m,d} = 1 - \frac{s_{m,d} - s_{\min}}{s_{\max} - s_{\min}}$$

Here, $s_{\min}$ and $s_{\max}$ represent the minimum and maximum $s_{m,d}$ values across all (m, d) pairs. A higher $\mathcal{S}_{m,d}$ indicates better alignment with ground truth, i.e., fewer and/or less severe discrepancies per utterance.

This severity-informed metric provides a principled way to compare trace fidelity and model reasoning performance across dialogue complexities.

4.3 Human Validation of Discrepancy Identification via LLM Judge

To examine judge reliability, we manually validated discrepancy identification on Dialogue 1. For all

	Dialogues					
	D1	D2	D3	D4	D5	D6
Annotator						
o3-mini						
Belief Contradictions	30	27	51	15	24	34
Omissions	55	49	101	52	88	49
Unsupported Beliefs	86	41	79	158	156	125
False Beliefs	6	3	3	0	6	0
Total Discrepancies (o3-mini)	204	**120**	234	225	274	**208**
Claude Sonnet 4						
Belief Contradictions	149	139	206	106	178	153
Omissions	144	115	226	76	209	160
Unsupported Beliefs	231	179	305	262	303	230
False Beliefs	4	2	8	0	7	4
Total Discrepancies (Claude)	**530**	**435**	**745**	**444**	**697**	**547**
Gemma 8.5B						
Belief Contradictions	121	114	191	103	149	135
Omissions	1	0	0	0	0	0
Unsupported Beliefs	62	49	101	53	96	85
False Beliefs	2	0	0	1	2	0
Total Discrepancies (Gemma)	**186**	163	292	157	**247**	220
Human						
Belief Contradictions	127	198	181	37	227	188
Omissions	52	20	28	52	128	103
Unsupported Beliefs	8	4	13	52	88	68
False Beliefs	0	0	3	0	0	0
Total Discrepancies (Human)	188	222	**225**	**141**	443	359

Table 2: LLM Discrepancy analysis results across 6 dialogues for each of 4 discrepancy types for all four LLMs. Each row indicates the number of a discrepancy type identified for a given dialogue. An LLM judge compared each model's audio-only mental model trace against the ground truth trace; naive human audio-only traces were evaluated using Claude Sonnet 4 as judge. Blue numbers highlight the lowest total number of discrepancies in a dialogue across all models and traces and purple numbers highlight the highest total discrepancies in a certain dialogue across all models and traces.

LLM trace comparisons, the same model that generated the audio-only trace served as the LLM judge; for naive human trace comparisons, Claude Sonnet 4 was used as judge. Table 3 reports overall accuracy per judge on Dialogue 1.

Judge	Correct	Wrong	Accuracy
Claude Sonnet 4	383	145	0.725
Gemma 8.5B	147	39	0.791
Naive Human	98	89	0.524
o3-mini	155	22	**0.876**

Table 3: Number of correct and incorrect discrepancy identifications for Dialogue 1.

5 Results

We report discrepancy counts across four categories: *belief contradictions*, *omissions*, *unsupported beliefs*, and *false beliefs* for three LLMs and a naive human baseline, each evaluated against ground-truth trace across six situated CReST dialogues.

5.1 Discrepancy Trends

Table 2, Table 4, and Figure 4 summarize total discrepancies per model's audio-only trace across

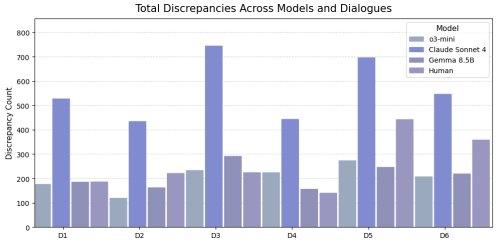


Figure 4: Visualization of the total discrepancies identified between each model's audio-only mental model trace and the ground truth trace. It is clear that Sonnet 4 identifies a large number more than other models, which as we discuss in Results, could be a positive or negative depending on how it is used. Visualizations of specific discrepancy types may be found in Appendix D.

all six dialogues. Across both a basic counting metric and normalized weighted score, we see the following. **Claude Sonnet 4** exhibits the highest overall discrepancy counts, surpassing 700 discrepancies in Dialogue 3 and having consistently the lowest discrepancy score across all dialogues as shown in Table 4. The model frequently introduces *unsupported beliefs* and *belief contradictions*, suggesting a tendency toward overgeneralized or speculative reasoning. **o3-mini** and **Gemma 8.5B** produce markedly fewer discrepancies, but with distinct profiles. o3-mini displays high *omission* rates and numerous *unsupported beliefs*, while Gemma identifies substantial *belief contradictions*. False beliefs are notably sparse across all models and traces (Table 6), which we attribute to the open-ended trace generation format: outright factual errors are less likely when audio-only traces record no change rather than commit to an incorrect belief. **Naive human annotators** demonstrate a discrepancy distribution that overlaps considerably with LLM profiles, particularly in *belief contradictions* and *omissions*, but outperform LLMs in limiting unsupported beliefs.

Across all models, certain dialogues (D3 and D5) consistently elicit higher discrepancy counts, indicating increased tracking difficulty, likely due to their longer lengths but possibly also conversational friction (e.g., (Sarkar et al., 2025b)). To validate that discrepancy differences across models are statistically significant, we conducted a two-way ANOVA across 519 model-utterance observations (3 LLM MM trackers x 1,142 utterances, reduced to matched observations across dialogues) (see Appendix C). Results confirm a strong main effect of model ($F(2,516)=99.82$, $p<\$2e-16$)

with no meaningful effect of utterance position (F(172,346)=0.593, p=0.9999), and pairwise comparisons show Claude differs significantly from both Gemma and o3-mini (both p <$2e-16), while Gemma and o3-mini do not differ (p=1.00).

Dialogue	o3-mini	Claude Sonnet 4	Gemma 8.5B	Human
D1	**0.917**	0.104	0.896	0.894
D2	**1.000**	0.141	0.883	0.722
D3	0.910	0.000	0.807	**0.926**
D4	0.770	0.229	0.939	**0.978**
D5	0.871	0.168	**0.916**	0.590
D6	**0.897**	0.197	0.873	0.585

Table 4: Summary of normalized discrepancy scores $\mathcal{S}_{m,d}$ for each model across six dialogues with w_x set to 1 for all weights. Bolded values represent the highest for each dialogue. Appendix E shows more discrepancy specifics.

5.2 Comparison to Human Audio-Only Traces

Naive human audio-only traces contain fewer unsupported beliefs than any LLM, but comparable or greater numbers of belief contradictions and omissions. This suggests that, while humans are better at avoiding the LLMs' hallucinated or unjustified inferences, they are also vulnerable to erroneous inferences about task state or agent perspective when lacking access to environmental context.

In terms of discrepancy counts, the model whose profile most closely matches human audio-only traces is **Gemma 8.5B**. However, this similarity has a trade-off: while Gemma avoids omissions, it more frequently makes incorrect inferences about ground truth belief states. We note that per-utterance normalization partially mitigates, but does not eliminate, differences in trace verbosity across models; a model that generates fewer field updates by default may incur fewer omissions by design.

5.3 Implications for Situated Theory of Mind

These findings highlight the nuanced but limited capability of current LLMs to perform ToM-like inference in grounded, task-based dialogue. While LLMs can maintain internally consistent belief structures across dialogue turns, they frequently fail to ground those beliefs in environmental constraints. The prevalence of unsupported beliefs and contradictions in Claude suggests limitations in evidence prioritization and contextual anchoring, while the divergence between o3-mini's high omission rate and Gemma's near-zero omissions points to fundamentally different trace generation strategies, underspecification versus overcommitment, across model families.

While LLMs demonstrate surface-level ToM capabilities in terms of belief attribution and second-order mental state modeling, their performance remains qualitatively distinct from that of humans. They are designed to be driven more by linguistic regularity than grounded situational reasoning.

6 Discussion & Conclusion

This study highlights a key limitation of current LLMs: while they can simulate mental state attribution to some extent, the lack of grounded situational reasoning often resulted in incorrect assessments of discrepancies between mental models. While our results show that LLMs can produce internally consistent mental model traces, they often fail to anchor these in environmental context, particularly in spatial reasoning scenarios. This gap reflects a broader limitation: LLMs infer mental states from linguistic patterns rather than from integrated spatial, temporal, and epistemic understanding. In our framework, unsupported beliefs most clearly illustrate this limitation. LLM annotators frequently introduce plausible but ungrounded inferences, especially in spatially ambiguous or disfluent dialogue turns. Because both annotators and judges operate on text-only transcripts, these errors also reflect the inability to resolve ambiguities that would otherwise be clarified through visual or prosodic cues. Overall, while LLMs exhibit measurable ToM-like behavior in situated dialogue, their reasoning remains fundamentally limited by a lack of grounding. Improving performance in this setting will likely require incorporating structured environmental context and multimodal inputs to better support accurate mental state tracking.

7 Limitations

Our experiments have several limitations. First, our results depend on the behavior of large language models, which can introduce variability due to non-deterministic outputs, prompt sensitivity, and ongoing model updates. Although we evaluate multiple models from different providers and observe consistent trends across dialogues, some variability in traces and discrepancies may still arise from prompt phrasing and model-specific behavior.

Second, while we include multiple models, our evaluation is not exhaustive. Differences observed across models suggest that discrepancy patterns

may vary with model architecture and training data.

Third, our experimental setup is constrained by the scale of the CReST corpus subset, which includes six dialogues with associated videos. This limited sample size restricts the diversity of interaction patterns and may impact the robustness of our findings. Expanding to larger and more varied dialogue datasets would strengthen the generality of the results.

Finally, our framework operates purely on text-based traces without incorporating structured environmental context. In spatially grounded tasks, providing LLM trace generators with explicit spatial representations (e.g., maps or scene layouts) or using VLMs may reduce unsupported beliefs and improve trace accuracy.[4] Exploring such grounded extensions is a promising direction for future work.

8 Ethical Considerations

In selecting models for this study, we prioritized a balance between experimental rigor and environmental responsibility. Large-scale evaluation across dozens of LLMs would have significantly increased compute time and energy consumption, contributing to the growing carbon footprint associated with LLM research. Instead, we chose a representative sample of three high-performing models from major LLM paradigms: OpenAI's o3-mini (proprietary autoregressive transformer), Claude Sonnet 4(Constitutional AI and safety-aligned), and Gemma (locally run and smaller at 8.5B parameters). These selections reflect a diverse cross-section of architectures, alignment strategies, and interface styles, providing meaningful comparative insights without unnecessary ecological impact. It is the belief of the authors that there is not a need to exhaust environmental resources when in fact our results show similar trends across even these three vastly different models. We encourage future work to adopt similar principled selection criteria to mitigate environmental harm while still advancing scientific understanding.

Acknowledgements

This work was in part funded by DARPA contract #HR001124C0502.

[4]We specifically did not utilize VLMs in this initial investigation to avoid any additional complications resulting from errors in image interpretation.

References

Angus Addlesee, Weronika Sieińska, Nancie Gunson, Daniel Hernández Garcia, Christian Dondrup, and Oliver Lemon. 2023. Multi-party goal tracking with llms: Comparing pre-training, fine-tuning, and prompt engineering.

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.

Leonardo Bertolazzi, Albert Gatt, and Raffaella Bernardi. 2024. A systematic analysis of large language models as soft reasoners: The case of syllogistic inferences. *Preprint*, arXiv:2406.11341.

Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, and Minlie Huang. 2024. Tombench: Benchmarking theory of mind in large language models. *Preprint*, arXiv:2402.15052.

Kathleen M Eberhard, Hannele Nicholson, Sandra Kübler, Susan Gundersen, and Matthias Scheutz. 2010. The "indiana cooperative remote search task"(CReST) corpus. In *LREC*.

Gwendolyn Edgar, Ayca Aygun, Matthew McWilliams, and Matthias Scheutz. 2024. Toward genuine robot teammates: Improving human-robot team performance beyond shared mental models with proactivity. In *Discovering the Frontiers of Human-Robot Interaction: Insights and Innovations in Collaboration, Communication, and Control*, pages 1–22. Springer.

Kanishk Gandhi, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah Goodman. 2023. Understanding social reasoning in language models with language models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Felix Gervits, Kathleen Eberhard, and Matthias Scheutz. 2016. Team communication as a collaborative process. *Frontiers in Robotics and AI*, 3:62.

P.N. Johnson-Laird. 1980. Mental models in cognitive science. *Cognitive Science*, 4(1):71–115.

Catholijn M Jonker, M Birna Van Riemsdijk, and Bas Vermeulen. 2010. Shared mental models: A conceptual analysis. In *International workshop on coordination, organizations, institutions, and norms in agent systems*, pages 132–151. Springer.

Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. 2023. Fantom: A benchmark for stress-testing machine theory of mind in interactions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, page 14397–14413, Singapore. Association for Computational Linguistics.

Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, Xin Zhou, Jiaming Zhou, and Haoqin Sun. 2024. Self-prompt tuning: Enable autonomous role-playing in llms.

Michal Kosinski. 2024. Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, 121(45):e2405460121. ArXiv:2302.02083 [cs].

Sandra Kübler, Eric Baucom, and Matthias Scheutz. 2012. Parallel syntactic annotation in crest. *Linguistic Issues in Language Technology*, 7.

Jiawen Liu, Yuanyuan Yao, Pengcheng An, and Qi Wang. 2024a. Peergpt: Probing the roles of llm-based peer agents as team moderators and participants in children's collaborative learning. *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, page 1–6.

Zijun Liu, Yanzhe Zhang, Peng Li, Yang Liu, and Diyi Yang. 2024b. A dynamic llm-powered agent network for task-oriented agent collaboration. (arXiv:2310.02170). ArXiv:2310.02170.

Ziqiao Ma, Jacob Sansom, Run Peng, and Joyce Chai. 2023. Towards a holistic landscape of situated theory of mind in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*.

Kumar Manas, Stefan Zwicklbauer, and Adrian Paschke. 2024. Tr2mtl: Llm based framework for metric temporal logic formalization of traffic rules. In *2024 IEEE Intelligent Vehicles Symposium (IV)*, pages 1206–1213.

John E Mathieu, Tonia S Heffner, Gerald F Goodwin, Eduardo Salas, and Janis A Cannon-Bowers. 2000. The influence of shared mental models on team process and performance. *Journal of applied psychology*, 85(2):273.

Philipp Mondorf and Barbara Plank. 2024. Beyond accuracy: Evaluating the reasoning behavior of large language models – a survey. *Preprint*, arXiv:2404.01869.

Amie Paige, Adil Soubki, John Murzaku, Owen Rambow, and Susan E. Brennan. 2024. Training llms to recognize hedges in dialogues about roadrunner cartoons. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, page 204–215, Kyoto, Japan. Association for Computational Linguistics.

D Palmer, Y Zhu, K Lai, H VanderHoeven, M Bradford, I Khebour, C Mabrey, J Fitzgerald, N Krishnaswamy, M Palmer, et al. Speech is not enough: interpreting nonverbal indicators of common knowledge and engagement (2024).

Liangming Pan, Alon Albalak, Xinyi Wang, and William Yang Wang. 2023. Logic-lm: Empowering large language models with symbolic solvers for faithful logical reasoning. *arXiv preprint arXiv:2305.12295*.

Nicholas Pangakis and Samuel Wolken. 2024. Knowledge distillation in automated annotation: Supervised text classification with llm-generated training labels.

Rupak Sarkar, Neha Srikanth, Taylor Pellegrin, Rachel Rudinger, Claire Bonial, and Philip Resnik. 2025a. Understanding common ground misalignment in goal-oriented dialog: A case-study with ubuntu chat logs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3200–3215.

Rupak Sarkar, Neha Srikanth, Taylor Pellegrin, Rachel Rudinger, Claire Bonial, and Philip Resnik. 2025b. Understanding common ground misalignment in goal-oriented dialog: A case-study with ubuntu chat logs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3200–3215, Vienna, Austria. Association for Computational Linguistics.

Karahan Sarıtaş, Kıvanç Tezören, and Yavuz Durmazkeser. 2025. A systematic review on the evaluation of large language models in theory of mind tasks. *Preprint*, arXiv:2502.08796.

Beau G Schelble, Christopher Flathmann, Nathan J McNeese, Guo Freeman, and Rohit Mallick. 2022. Let's think together! assessing shared mental models, performance, and trust in human-agent teams. *Proceedings of the ACM on Human-Computer Interaction*, 6(GROUP):1–29.

Matthias Scheutz, Scott A DeLoach, and Julie A Adams. 2017. A framework for developing and using shared mental models in human-agent teams. *Journal of Cognitive Engineering and Decision Making*, 11(3):203–224.

Matthias Scheutz, Bradley Oosterveld, John Peterson, Eric Wyss, and Evan Krause. 2024. A multi-robot architecture framework for effective robot teammates in mixed-initiative teams. In *Proceedings of the 2024 International Symposium on Technological Advances in Human-Robot Interaction*, pages 74–82.

Parshin Shojaee*†, Iman Mirzadeh*, Keivan Alizadeh, Maxwell Horton, Samy Bengio, and Mehrdad Farajtabar. 2025. The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity.

Kaya Stechly, Karthik Valmeekam, and Subbarao Kambhampati. 2024. On the self-verification limitations of large language models on reasoning and planning tasks. *Preprint*, arXiv:2402.08115.

James WA Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, et al. 2024. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 8(7):1285–1295.

Tomer Ullman. 2023. Large language models fail on trivial alterations to theory-of-mind tasks. (arXiv:2302.08399). ArXiv:2302.08399 [cs].

Leonie Weissweiler, Abdullatif Köksal, and Hinrich Schütze. 2024. Hybrid human-llm corpus construction and llm evaluation for rare linguistic phenomena.

Pengying Wu, Yao Mu, Kangjie Zhou, Ji Ma, Junting Chen, and Chang Liu. 2024. Camon: Cooperative agents for multi-object navigation with llm-based conversations.

Danni Yu, Luyang Li, Hang Su, and Matteo Fuoli. 2024. Assessing the potential of ai-assisted pragmatic annotation: The case of apologies. (arXiv:2305.08339). ArXiv:2305.08339.

Ranchi Zhao, Zhen Leng Thai, Yifan Zhang, Shengding Hu, Yunqi Ba, Jie Zhou, Jie Cai, Zhiyuan Liu, and Maosong Sun. 2024. Decoratelm: Data engineering through corpus rating, tagging, and editing with language models. *arXiv preprint arXiv:2410.05639*.

Jin Peng Zhou, Charles Staats, Wenda Li, Christian Szegedy, Kilian Q Weinberger, and Yuhuai Wu. 2024. Don't trust: Verify–grounding llm quantitative reasoning with autoformalization. *arXiv preprint arXiv:2403.18120*.

Yifan Zhu, Kenneth Lai, Ibrahim Khebour, Marc Verhagen, Nikhil Krishnaswamy, and James Pustejovsky. 2025. Multimodal situational awareness: Neurosymbolic ai for real-time hci in the classroom. In *International Conference on Human-Computer Interaction*, pages 454–464. Springer.

A Full Prompts

The following section outlines our prompts used for the experiments. These prompts were all given the exact same way to all models to compare models as effectively as possible. To note, we tried many different prompts during this project including persona prompting, chain of thought prompting, and more. Across initial experiments we determined that these prompts performed the best and therefore we ran primary experiments with them. Note: The prompt below uses the term 'annotator belief' as a shorthand label within the JSON output schema. In our framework, this corresponds to the belief recorded in the audio-only mental model trace, as described in Section 4.

Annotation Prompt

```
You are an annotator of a dialogue of a 2-person team working together on a search mission.
You will be provided with several rules and examples of how to perform belief, commitment, and goal updates based on dialogue history.
You are only tracking the beliefs, goals, and commitments which share information related to the searcher's location and level of
task completion.

TEAM MEMBERS:
1. Searcher: The searcher is in a remotely located real world environment of a floor in a building with many rooms, boxes, and doors.
2. Director: The director is local but has a static (but somewhat flawed) map of the floor showing some of the boxes, rooms and doors,
and is communicating and coordinating via audio with the searcher.

YOUR TASK:
You will carefully listen to the conversation (current dialogue move and recent dialogue history) and logically deduce/induce/abduce
beliefs, goals, and commitments held by the director and searcher suggested by the dialogue move in context of the dialogue history.
You will be given a prior state of the world and your task is to decide what beliefs, goals, and commitments to add and what to remove
to this state based on what was said in the dialogue move and history.
You have to deduce the beliefs, goals, and commitments yourself. You should be thinking from the perspective of the searcher when
determining their beliefs, goals, and commitments and vice versa for the director.
This does not mean that the searcher and director beliefs will align.

For each utterance, you must return a JSON object with the following structure:
{
    "speaker": <speaker>,
    "utterance": <utterance>,
    "start": <start>,
    "end": <end>,
    "Annotation": {
        "Searcher believes": <string>,
        "Director believes": <string>,
        "2nd order: Searcher believes that the director believes": <string>,
        "2nd order: Director believes that the searcher believes": <string>,
        "Searcher has committed to": <string>,
        "Director has committed to": <string>,
        "Director's goal is": <string>,
        "Searcher's goal is": <string>,
        "Common Belief": <string>
    }
}
For any field where you do not think there is an update, write "no change".
For the "Common Belief" field, if there is a clear belief that both agents share, summarize it in 1-2 sentences. Otherwise, write
"No change".

WHAT ARE BELIEFS, GOALS, AND COMMITMENTS?
Beliefs: the understanding and assumptions that an agent holds about a situation, task, or team dynamic I.E. where an agent is,
what they're doing, the state of the world
Goals: the objectives or desired outcomes that an agent strives to achieve
Commitment: agreement of an agent to their individual responsibilities, actions, and expected contributions towards achieving
a goal.

EXAMPLES OF HOW TO RESPOND:

EXAMPLE 1:

DIALOGUE HISTORY:
Searcher: "So we're supposed to get the green boxes?"
Director: "I think so."

CURRENT DIALOGUE MOVE:
Searcher: "Okay."

OUTPUT:
{
    "speaker": "Searcher",
    "utterance": "Okay.",
    "start": <start>,
    "end": <end>,
    "Annotation": {
        "Searcher believes": "no change",
        "Director believes": "no change",
        "2nd order: Searcher believes that the director believes": "no change",
```

```
        "2nd order: Director believes that the searcher believes": "no change",
        "Searcher has committed to": "no change",
        "Director has committed to": "no change",
        "Director's goal is": "no change",
        "Searcher's goal is": "The searcher's goal is to get the green boxes.",
        "Common Belief": "Both the searcher and director have the shared goal to get green boxes."
    }
}

EXAMPLE 2:

DIALOGUE HISTORY:
Director: "okay . in front of your there should be a platform with steps going  up?"
Searcher: "right"
Director: "okay so make a right turn"
Searcher: "kay"
Director: "okay a:nd walk into the next room . through the- there should be an open door there"
Searcher: "right"
Director: "okay walk into the next room"
Searcher: "kay"
Director: "okay . now you should be in a roo:m wi:th . like on the right there should be like two cubicles"
Searcher: "yea there's two like two little rooms"
Director: "ah-"
Director: "okay and straight in front of you should be: . filing cabinet?"

CURRENT DIALOGUE MOVE:
Searcher: "yes"

OUTPUT:
{
    "speaker": "Searcher",
    "utterance": "yes",
    "start": <start>,
    "end": <end>,
    "Annotation": {
        "Searcher believes": "The searcher believes that there is a filing cabinet in front of them.",
        "Director believes": "no change",
        "2nd order: Searcher believes that the director believes": "The searcher believes that the director believes there is a
        filing cabinet in front of them.",
       "2nd order: Director believes that the searcher believes": "The director believes the searcher believes there is a filing
        cabinet in front of the searcher.",
        "Searcher has committed to": "no change",
        "Director has committed to": "no change",
        "Director's goal is": "no change",
        "Searcher's goal is": "no change",
        "Common Belief": "Both agents believe there is a filing cabinet in front of the searcher."
    }
}

Learn from these examples and generalize to new cases as you reason through the dialogue you are given.
You need to infer new states from the past ones. Think from the perspectives of both the director and searcher and what
they're thinking at each time step.

The point of these annotations is to maintain an understanding of the shared mental models of the searcher and director.
We want to track what the searcher and director each independently think is going on in the task based on their perception
of the searcher's location.
This allows us to investigate the coherence of the team.

You should only use the following verbs when trying to identify a belief, goal, or commitment:
    at,
    in,
    holding,
    connects,
    near,
    right of,
    in front of,
    on,
    across from,
    get,
    go,
    turn,
    find

TAKE THESE STEPS TO PERFORM YOUR TASK ON EACH MOVE:
1. Identify the speaker of the current dialogue move.
2. Identify the dialogue act of the move. Could be one of question | assertion/statement | command/request | offer | promise |
acknowledgement | greeting. This can influence what you conclude about the beliefs. For example, questions about p suggest
that the speaker did not previously believe p.
3. Identify the current state of the world.
4. Based on the current dialogue move and the history, describe what beliefs, goals, and commitments of the director and
searcher you can infer (implicitly or explicitly). Use ONLY the phrasing "The [searcher/director] believes", "The [searcher/director]
is committed to", "The [searcher/director]'s goal is" to begin the annotation.
5. Reason through this and rationalize why you included the updates that you chose to include.

If no updates are needed, write "no change" for every field in the Annotation object.

Output MUST be a JSON string, and NOT markdown.
```

```
IMPORTANT: Output ONLY the JSON object, with no explanation or commentary. Do not include any reasoning or
markdown formatting.
```

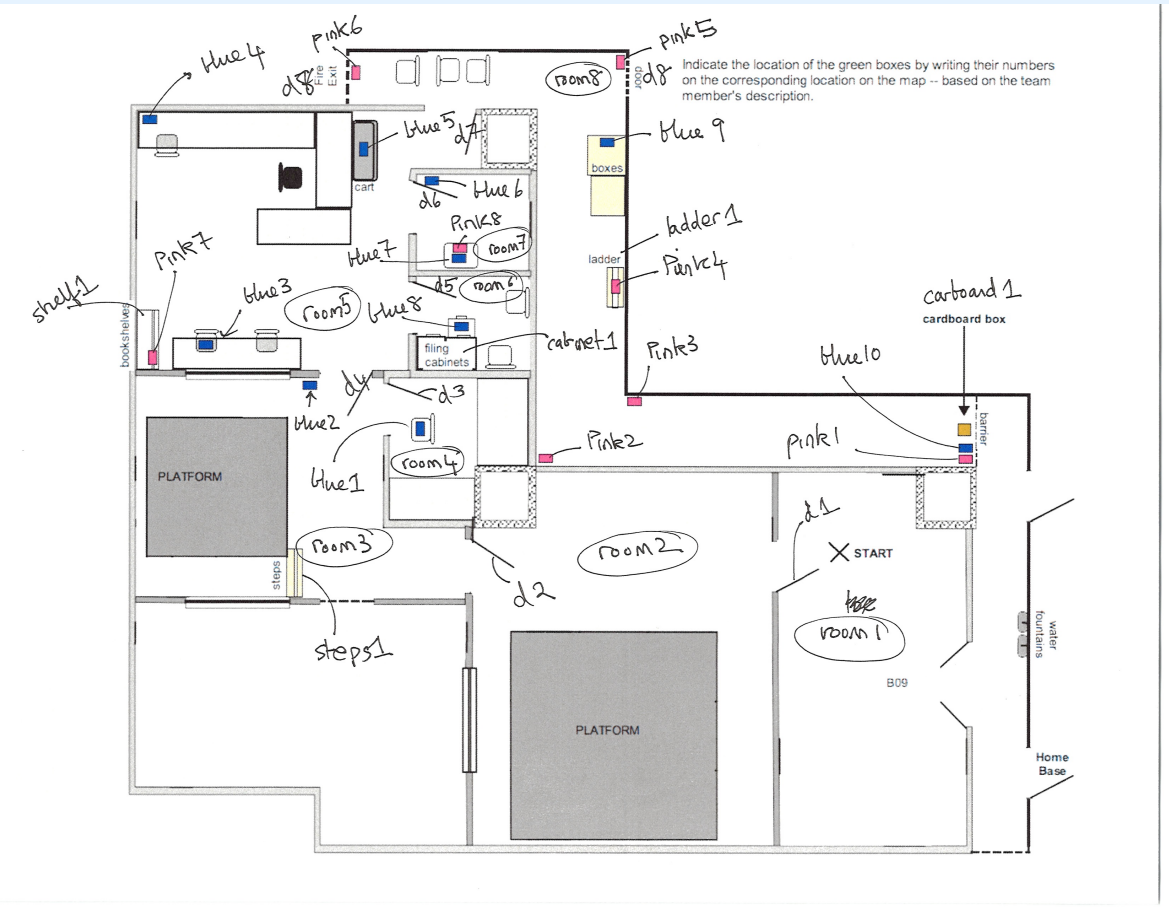
Discrepancy Detection Prompt

```
You are given two sets of annotations representing the Shared Mental Models (SMMs) of two agents in a dialogue. Your task is to
identify and classify discrepancies between the two agents' mental models using the following five types:

1. Belief Contradiction: One annotator identifies a belief b, and the other identifies not b.
2. False Belief: An annotator believes something that contradicts the known ground truth.
3. Omission: The ground truth includes a belief that the annotator simply omits.
4. Unsupported Belief: A belief not verifiable from the context or ground truth; lacks support either way.

For each discrepancy:
- Classify it into one or more of the types above.
- Clearly state what each agent believes, using plain English.
- Provide a short explanation.

Output Format:
You must output a single JSON object with the following structure. For each discrepancy found, create an object with the keys:
'Discrepancy Type', 'Ground Truth Belief', 'Annotator Belief', and 'Explanation'. If no discrepancies are found, return an empty array.

Required JSON structure:
    "Discrepancies": [
        {
            "Discrepancy Type": "Belief Contradiction",
            "Ground Truth Belief": "belief from ground truth",
            "Annotator Belief": "belief from annotator",
            "Explanation": "explanation of the discrepancy"
        }
    ]

Examples:

Example 1: Belief Contradiction
{
    "Discrepancies": [
        {
            "Discrepancy Type": "Belief Contradiction",
            "Ground Truth Belief": "The searcher is at the cardboard box",
            "Annotator Belief": "The searcher is at room 1",
         "Explanation": "The ground truth indicates the searcher is at the cardboard box, while the annotator believes the searcher
            is at room 1. These beliefs are directly contradictory."
        }
    ]
}

Example 2: False Belief
{
    "Discrepancies": [
        {
            "Discrepancy Type": "False Belief",
            "Ground Truth Belief": "The searcher is in room r",
            "Annotator Belief": "The searcher is not in room r",
          "Explanation": "The annotator believes the searcher is not in room r, but the ground truth indicates that the searcher is
            actually there."
        }
    ]
}

Example 3: Omission
{
    "Discrepancies": [
        {
            "Discrepancy Type": "Omission",
            "Ground Truth Belief": "The director's goal is for the searcher to go to corridor 2",
            "Annotator Belief": "No mention of director's goal",
           "Explanation": "The ground truth includes a goal for the searcher to go to corridor 2, while the annotator omits this
            information entirely."
        }
    ]
}

Example 4: Unsupported Belief
{
    "Discrepancies": [
        {
            "Discrepancy Type": "Unsupported Belief",
            "Ground Truth Belief": "No specific belief about other agent's thoughts",
            "Annotator Belief": "The annotator believes that the other agent thinks the searcher is in room 1",
          "Explanation": "The annotator makes an assumption about the other agent's belief that cannot be verified from the given
            dialogue or any known facts."
        }
    ]
}
```

```
Example 5: No Discrepancies
{
    "Discrepancies": []
}
```

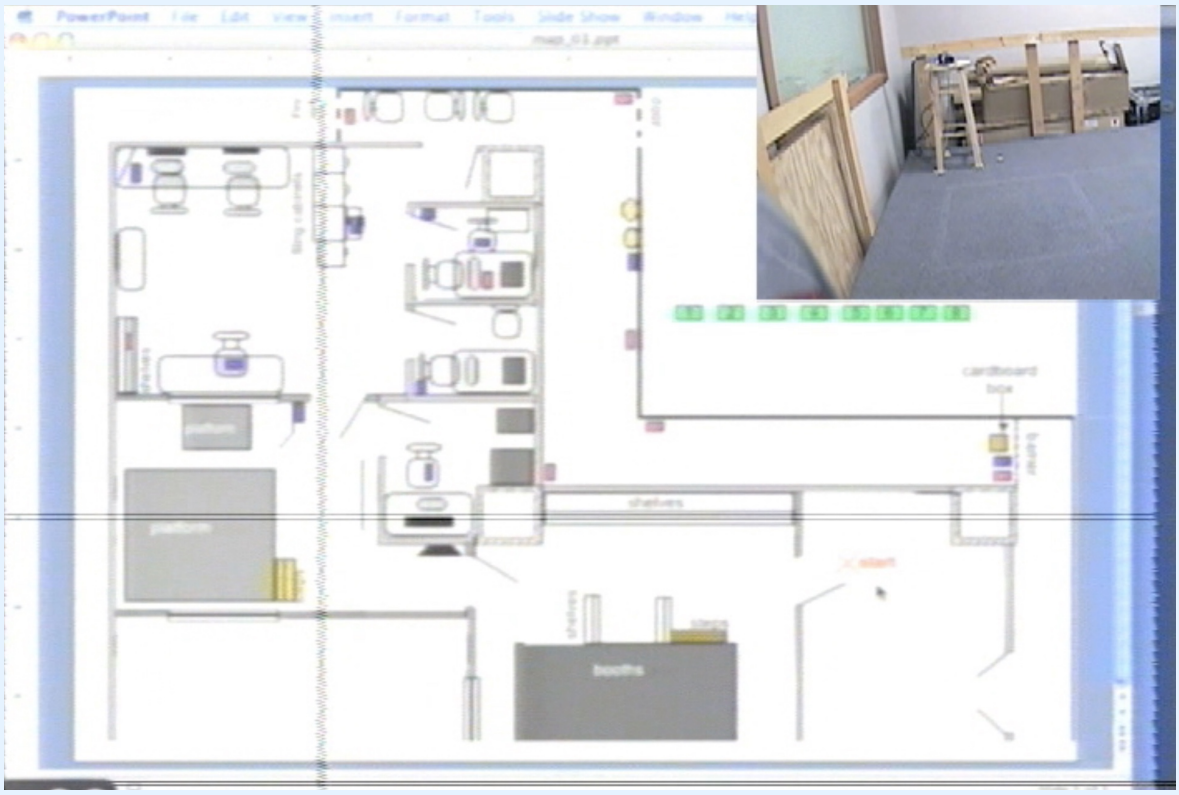
B CReST Dataset Information

Map of Environment



The map supplied to the director in the CReST experiments, which is flawed in some minor ways so as to induce confusion between the searcher and director. The director guides the searcher through this environment to the blue boxes as the searcher scans for green boxes, and this map's locations of blue boxes is partially incorrect. Here, we labeled the locations of objects and rooms as given to the LLM, though that was not given to the director at the start of the experiment. The searcher is in this environment and is not given this information.

Video Screenshot of Environment



Here we have supplied a screenshot from a video of the CReST corpus. Videos such as this one were used to create the dataset of ground truth mental model traces for each dialogue, which was corroborated by two human annotators. Here, one can see that there is an image of the map in the director's view and the black lines forming a cross show the searcher's current location, which dynamically moves during the video to update the real-time searcher's location. In the top right hand corner is the headcam video from the searcher's perspective, which shows exactly what they are looking at. Because of these videos, we were able to establish a ground truth understanding of the environment external to the naive human and LLM trace generators, which we use in our evaluations of the LLM discrepancy finder.

C Statistical Validation of Results

To validate the statistical significance of this study, we ran ANOVA tests between the two conditions of naive human and LLM audio-only traces, as evaluated by the LLM judge. Our two-way ANOVA is pasted below, with the p-value of less than 2e-16.

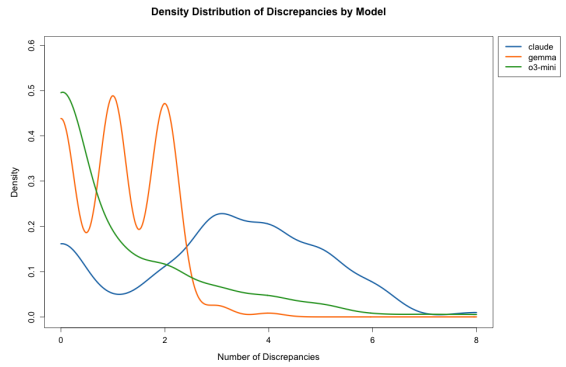
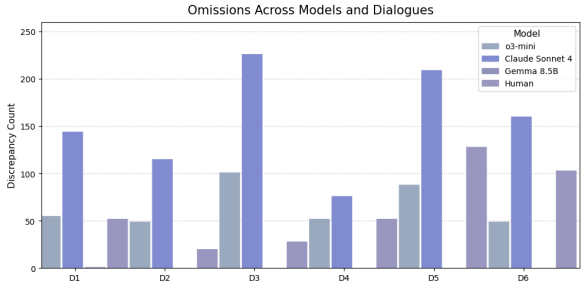


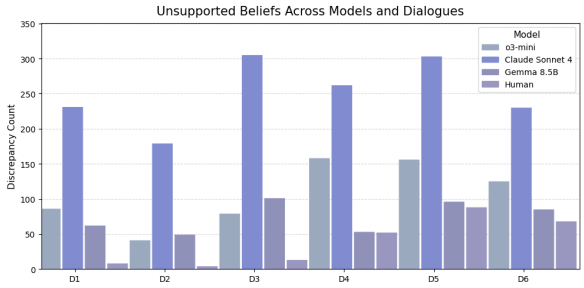
Figure 5: Violin plot of discrepancy density across LLMs-As-A-Judge.

Most notably, we found a model effect where LLM judges showed a difference in identified discrepancies. Across 519 model-utterance observations, the violin plot shows a clear separation in discrepancy distributions by model: Claude is centered at substantially higher discrepancy counts and has a broader spread, whereas gemma and o3-mini are concentrated at low counts with strong overlap (both near 1 discrepancy on average). This visual pattern is confirmed by ANOVA, which found a strong main effect of model on discrepancies ($F(2,516)=99.82$, $p<2e\text{-}16$), but no meaningful effect of utterance number ($F(172,346)=0.593$, $p=0.9999$). Pairwise comparisons (Bonferroni-adjusted) further indicate that Claude differs significantly from both Gemma and o3-mini (both $p<2e\text{-}16$), while gemma and o3-mini do not differ ($p=1.00$), supporting the conclusion that discrepancy burden is primarily model-driven rather than position-in-dialogue-driven.

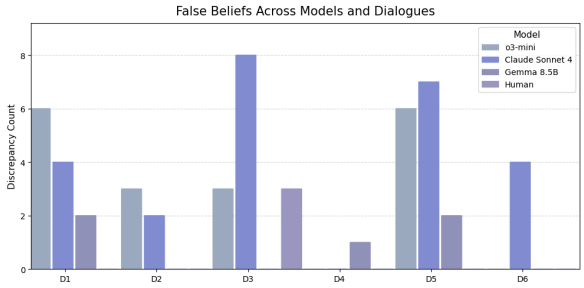
D Discrepancy Visualizations



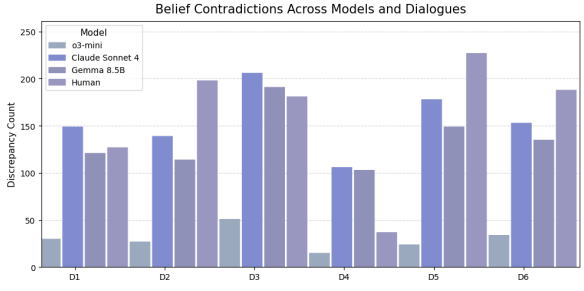
This figure shows the number of omissions identified by each LLM across the dialogue corpus.



This figure visualizes the number of unsupported beliefs returned by each LLM.



This figure displays false belief discrepancies identified when comparing LLM audio-only traces against the ground truth trace.



This figure shows contradictions between ground truth and LLM-derived beliefs.

E Discrepancy Details

These tables display, for each discrepancy type, the raw count of errors divided by the number of utterances in each dialogue, yielding a fair measure of how often that error occurs per turn. Presenting each discrepancy type in its own table makes it easy to spot patterns in belief contradictions, false beliefs, unsupported beliefs, and omissions before any further scaling.

Dialogue	o3-mini	Claude Sonnet 4	Gemma 8.5B	Naive Human
D1	0.173	0.861	0.699	0.734
D2	0.184	0.946	0.776	1.347
D3	0.227	0.916	0.849	0.804
D4	0.093	0.654	0.636	0.228
D5	0.100	0.739	0.618	0.942
D6	0.175	0.789	0.696	0.969

Table 5: Per-utterance rates for Belief Contradictions

Dialogue	o3-mini	Claude Sonnet 4	Gemma 8.5B	Naive Human
D1	0.035	0.023	0.012	0.000
D2	0.020	0.014	0.000	0.000
D3	0.013	0.036	0.000	0.013
D4	0.000	0.000	0.006	0.000
D5	0.025	0.029	0.008	0.000
D6	0.000	0.021	0.000	0.000

Table 6: Per-utterance rates for False Beliefs

Dialogue	o3-mini	Claude Sonnet 4	Gemma 8.5B	Naive Human
D1	0.497	1.335	0.358	0.046
D2	0.279	1.218	0.333	0.027
D3	0.351	1.356	0.449	0.058
D4	0.975	1.617	0.327	0.321
D5	0.647	1.257	0.398	0.365
D6	0.644	1.186	0.438	0.351

Table 7: Per-utterance rates for Unsupported Beliefs

Dialogue	o3-mini	Claude Sonnet 4	Gemma 8.5B	Naive Human
D1	0.318	0.832	0.006	0.301
D2	0.333	0.782	0.000	0.136
D3	0.449	1.004	0.000	0.124
D4	0.321	0.469	0.000	0.321
D5	0.365	0.867	0.000	0.531
D6	0.253	0.825	0.000	0.531

Table 8: Per-utterance rates for Omissions

These tables show an example of our metric applied to our dataset. They show the normalized scores for each model across a dialogue, taking into account the per-utterance scores above. This condenses all four discrepancy types into a single metric $\mathcal{S}_{m,d}$ for each model and dialogue, linearly scaled to the range [0, 1] so that higher values indicate cleaner performance. By dividing the weighted raw discrepancy counts by dialogue length and then normalizing against the global minimum and maximum scores, these tables provide a directly comparable measure of overall error rate per utterance. Viewing one table per model highlights which dialogues each system handles most accurately. This metric may be used with assigned weights, but for our use case they were all set to 1.

Dialogue	Normalized Score
D1	0.917
D2	1.000
D3	0.910
D4	0.770
D5	0.871
D6	0.897

Table 9: Normalized scores for `o3-mini`

Dialogue	Normalized Score
D1	0.104
D2	0.141
D3	0.000
D4	0.229
D5	0.168
D6	0.197

Table 10: Normalized scores for `Claude Sonnet 4`

Dialogue	Normalized Score
D1	0.896
D2	0.883
D3	0.807
D4	0.939
D5	0.916
D6	0.873

Table 11: Normalized scores for `Gemma 8.5B`

Dialogue	Normalized Score
D1	0.894
D2	0.722
D3	0.926
D4	0.978
D5	0.590
D6	0.585

Table 12: Normalized scores for `Naive Human`

F Qualitative Error Analysis

To substantiate our claims about spatial reasoning and disfluency failures, we present representative examples drawn from the human-validated discrepancy set for Dialogue 1. For each example we show the utterance, the ground truth trace, the LLM trace, the discrepancy type identified by the judge, and a brief explanation. Examples are selected to illustrate the three most common failure modes: spatial misgrounding, disfluency-induced hallucination, and correct trace for contrast.

Utterance	**Model**	**Discrepancy Type**	**Evidence**
"yes" (t=27.18s)	Gemma 8.5B	Unsupported Belief	The audio-only trace introduces a spatial second-order belief (that the searcher knows the platform and cubicle location) that is absent from the ground truth trace.
"and uh there's a cubicle on your right hand side and then straight in front of you there's also a door?" (t=22.45s)	o3-mini	Unsupported Belief	The audio-only trace introduces a director goal ('confirm environment features to navigate accurately') not present in the ground truth trace. The disfluency (uh) appears to trigger a spurious goal update.
"okay so u:m" (t=3.31s)	Claude Sonnet 4	None (correct)	No discrepancy flagged. LLM correctly withholds MM update for a turn consisting entirely of a filler token and elongated acknowledgment.

Table 13: Representative examples from the human-validated discrepancy set for Dialogue 1, illustrating a spatial misgrounding failure, a disfluency-induced unsupported belief, and a correct trace. These examples support the pattern described in Section 6: LLMs generate linguistically plausible but environmentally ungrounded inferences, particularly in turns where spatial reference or transcription artifacts are present.

These examples illustrate the pattern described in Section 6: LLMs generate linguistically plausible but environmentally ungrounded inferences, particularly in turns where spatial reference or transcription artifacts are present.