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

Katharine Kowalyshyn and Matthias Scheutz
Department of Computer Science, Human-Robot Interaction Laboratory, Tufts University, Medford, MA 02155, USA
{Katharine.Kowalyshyn, Matthias.Scheutz}@tufts.edu

Motivation and Research Question

Successful teams maintain a **shared mental model (SMM)**: aligned beliefs, goals, and commitments about the task. We ask whether LLMs can infer these latent states from **situated dialogue** and identify when their inferred mental models diverge from video-grounded human annotations.

CReST Situated Dialogue Corpus

- 23 dyads: a remote **director** with an incomplete map and an on-site **searcher**.
- Six dialogues with synchronized head-camera video establish environmental ground truth.
- **1,142 utterances** total across the selected dialogues.
- Tasks require navigation, box identification, and block relocation amid natural disfluencies.

Mental Model Trace

At each utterance, annotators update (1) first-order beliefs for searcher and director, (2) second-order beliefs (what each thinks the other believes), (3) goals and commitments and (4) common belief / shared task state.

Mental Model Structure

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

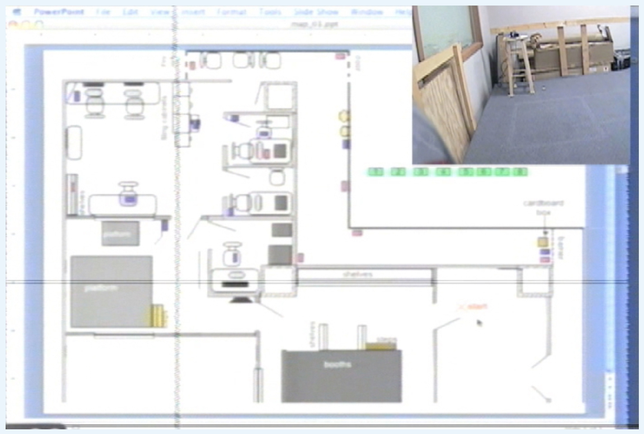
Current Dialogue Move:
Searcher: "Okay."

Current Mental Model:

```
"speaker": "Searcher",
  "utterance": "Okay.",
  "start": 5.41,
  "end": 8.753,
  "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."
```

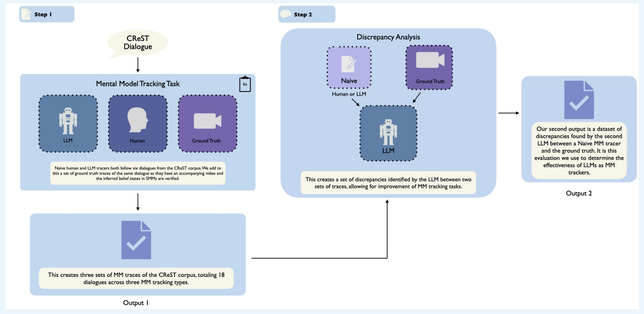
Each mental state update, regardless if it's tracked by an LLM or human, follows most of this structure. LLMs are requested to add the "Common Belief" field to identify the SMM in the current dialogue step for simplified analysis.

CReST Task



The director's map of the envioronment with the headcam view of the searcher (not shown to the director).

Two-Step Evaluation Pipeline



Step 1: Trace generation. Three LLMs, naive humans, and video-informed humans independently produce mental model traces.

Step 2: Discrepancy analysis. An LLM judge compares each audio-only trace against the video-grounded trace.

Models and Comparisons

- OpenAI **o3-mini**
- Anthropic **Claude Sonnet 4**
- Google **Gemma 8.5B**
- Naive human audio-only baseline

The experiment yields **30 trace sets** and **24 audio-only vs. ground-truth comparisons**.

A Novel Discrepancy Framework

A discrepancy is a field-level mismatch between an audio-only trace and the ground state determined by the video from the searcher's headcam.

Belief contradiction: Agents are assigned mutually exclusive beliefs about the same world state.

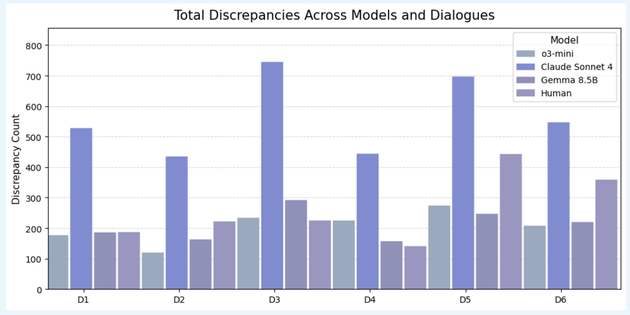
False belief: A proposition conflicts with the grounded state.

Unsupported belief: A plausible but ungrounded proposition is added.

Omission: A task-relevant update present in ground truth is missing.

Severity ranking: Contradictions > False beliefs > Unsupported beliefs > Omissions

Results: Total Discrepancies



- Claude Sonnet 4 produced the most discrepancies in every dialogue (up to **745** in D3), dominated by unsupported beliefs and contradictions.
- o3-mini and Gemma produced fewer total discrepancies, but with distinct failure profiles.
- Naive humans made very few unsupported-belief errors, but still showed contradictions and omissions without video context.

Judge Validation on Dialogue 1

Judge	Correct / Total	Accuracy
o3-mini	155 / 177	**0.876**
Gemma 8.5B	147 / 186	0.791
Claude Sonnet 4	383 / 528	0.725
Naive human	98 / 187	0.524

Key Findings

1. LLMs can generate internally coherent mental-state traces from dialogue.
2. Their coherence is often **surface-level**: beliefs are not reliably anchored to environmental constraints.
3. **Spatial ambiguity** and **disfluencies** systematically trigger errors.
4. Model families trade off **underspecification** (omissions) against **overcommitment** (unsupported or contradictory beliefs).
5. Human audio-only traces are also imperfect, showing the central role of perceptual grounding.

Statistical Evidence

A two-way ANOVA over matched model-utterance observations found a strong model effect: $F(2, 516) = 99.82$ **($p < 2 \times 10^{-16}$) with no meaningful effect of utterance position.**

Implications

Situated dialogue trace generation is a useful proxy for evaluating ToM-like behavior, but current text-only LLMs rely more on linguistic regularity than integrated spatial, temporal, and epistemic reasoning.

Conclusion and Future Work

Reliable mental-state tracking will likely require:

- structured maps or scene representations;
- multimodal or vision-language inputs;
- calibrated uncertainty rather than speculative updates;
- larger and more diverse grounded dialogue datasets.

Acknowledgements

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